UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

WALDYR LOURENÇO DE FREITAS JUNIOR

Um comparativo quantitativo e qualitativo de algoritmos de coagrupamento baseados em fatoração de matrizes

São Paulo
2023

WALDYR LOURENÇO DE FREITAS JUNIOR

Um comparativo quantitativo e qualitativo de algoritmos de coagrupamento baseados em fatoração de matrizes

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 23 de março de 2023. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientadora: Profa. Dra. Sarajane Marques Peres

São Paulo

2023

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Freitas Junior, Waldyr Lourenço de
Um comparativo quantitativo e qualitativo de algoritmos de coagrupamento baseados em fatoração de matrizes / Waldyr Lourenço de Freitas Junior; orientadora, Sarajane Marques Peres. -- São Paulo, 2023.
124 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2023.
Versão corrigida

1. Coagrupamento. 2. Fatoração de Matrizes. 3. Interpretação Humana. I. Peres, Sarajane Marques, orient. II. Título.

Dissertação de autoria de Waldyr Lourenço de Freitas Junior, sob o título **"Um comparativo quantitativo e qualitativo de algoritmos de coagrupamento baseados em fatoração de matrizes"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 23 de março de 2023 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Sarajane Marques Peres
Universidade de São Paulo
Presidente

Prof. Dr. Fabricio Olivetti de França
Universidade Federal do ABC

Profa. Dra. Rosana Retsos Signorelli Vargas
Universidade de São Paulo

À minha eterna namorada Francine por me apoiar incondicionalmente. Ao meu velhinho, Sr. Waldyr, que infelizmente não está mais ao meu lado; ele ficaria muito feliz por mim. À minha querida mãe, Sra. Nanete, por tantos anos de dedicação para cuidar de mim. Às minhas duas princesas, Yasmin e Lavínia, por muitas vezes terem aberto mão do tempo delas comigo, para eu me dedicar ao Mestrado. À minha orientadora Prof. Dra. Sarajane, pela paciência, instrução, amizade e orientação. À Universidade de São Paulo por me proporcionar educação de qualidade e a realização de um grande sonho.

"Educação não transforma o mundo.

Educação muda as pessoas.

Pessoas transformam o mundo."

(Paulo Freire)

Resumo

FREITAS JUNIOR, Waldyr Lourenço De. **Um comparativo quantitativo e qualitativo de algoritmos de coagrupamento baseados em fatoração de matrizes**. 2023. 124 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2023.

Agrupamento é uma estratégia para análise de dados que objetiva encontrar grupos cujos dados são mais similares entre si, enquanto dados organizados em grupos distintos são mais dissimilares entre si. Coagrupamento é uma estratégia semelhante, contudo, aplicado simultaneamente sobre os dados e atributos de um conjunto de dados. Diferentes contextos usam coagrupamento, tais como análise de imagens, bioinformática e mineração de textos. Para este último, cujos dados sob análise dizem respeito a contextos caracterizados por subjetividade, a literatura apresenta alguns poucos estudos relacionados à interação humana para interpretação dos resultados. Dentre uma série de abordagens de coagrupamento, destaca-se a fatoração tripla de matrizes não negativas (NMTF). Estudos reconhecem a utilidade dessa abordagem por seu alto desempenho e facilidade em trabalhar com dados diádicos e dados com alta dimensionalidade. Corpus textuais, cuja representação seja baseada no modelo de espaço vetorial, podem produzir matrizes de dados com alta dimensionalidade e alta esparsidade. Essas características tornam tais problemas candidatos a serem tratados por meio da abordagem NMTF. A literatura apresenta diferentes algoritmos de coagrupamento baseados em fatoração de matrizes; tais estudos concentraram-se na avaliação da capacidade de agrupamento dos algoritmos, mas não trataram o aspecto da qualidade dos resultados segundo a ótica da interpretação humana. Assim, o objetivo principal deste trabalho foi explorar sistematicamente um conjunto de algoritmos de coagrupamento baseados em fatoração de matrizes, com atenção à interpretação humana dos resultados produzidos por eles. Este trabalho também explorou esses algoritmos em diferentes circunstâncias e revelou mais claramente suas vantagens e desvantagens. Os experimentos se basearam em conjuntos de dados sintéticos e do mundo real. Os conjuntos de dados sintéticos foram rotulados e contavam com diferentes estruturas de cogrupos; o objetivo foi explorar a capacidade que os algoritmos têm em agrupar dados e atributos. Um conjunto de dados do mundo real usado como referência para tarefas de análise automática de textos foi escolhido para uso nos experimentos com dados do mundo real. O conjunto consiste de um corpus público de notícias (com e sem caráter de hiperpartidarismo), extraídas de diferentes sites entre 2016 e 2018; o objetivo foi realizar uma análise detalhada da robustez dos algoritmos sob uma análise qualitativa de resultados, realizada sob uma ótica de interpretação humana. Para essa análise qualitativa, foram realizadas uma série de tarefas baseadas em questionários estruturados aplicados a alunos de graduação da Universidade de São Paulo. Os experimentos com dados sintéticos e do mundo real demostraram que algoritmos com restrições binárias apresentam desempenho melhor que os demais. Além disso, uma análise de palavras que melhor representam grupos de notícias evidenciou dificuldades dos algoritmos em definir claramente, no sentido semântico, tais grupos. O algoritmo proposto neste trabalho (WC-FNMTF) foi submetido a diferentes tarefas e apresentou bons resultados. A tarefa com humanos revelou superioridade do algoritmo NBVD, seguido do WC-FNMTF.

Palavras-chaves: Coagrupamento. Fatoração de Matrizes. Interpretação Humana.

Abstract

FREITAS JUNIOR, Waldyr Lourenço De. **A quantitative and qualitative comparison of co-clustering algorithms based on matrix factorization**. 2023. 124 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2023.

Clustering is a strategy for data analysis to identify clusters whose data points are more similar to each other. Data points organized into distinct clusters are more dissimilar to each other. Co-clustering is a similar strategy, however, it is applied simultaneously to data and attributes of a data set. Different contexts use co-clustering, such as image analysis, bioinformatics, and text mining. For the latter, whose data under analysis concern contexts characterized by subjectivity, the literature presents a few studies related to human interaction for interpreting results. Among several co-clustering approaches, the Non-negative Matrix Factorization (NMTF) stands out. Studies recognize the usefulness of such an approach because of its high performance and ease of working with dyadic data and data with high dimensionality. Corpus, whose representation is based on the vector space model, can produce data matrices with high dimensionality and high sparsity. These characteristics make such problems candidates to be addressed through the NMTF approach. The literature presents different co-clustering algorithms based on matrix factorization; such studies focused on evaluating the algorithms' clustering ability but did not address quality aspects from the perspective of human interpretation of the meaning of the generated clusters. Thus, the main objective of this work was systematically to explore a set of co-clustering algorithms based on matrix factorization, with attention to human interpretation of the results produced by them. This work also explored such algorithms in different circumstances to reveal their advantages and disadvantages. Experiments were based on synthetic data sets and real-world data sets. The synthetic data sets were labeled and composed of different co-cluster structures; the goal was to explore algorithms' ability to cluster attributes and data. A real-world data set used as a reference for automatic text analysis tasks was chosen for experiments with real-world data. The data set comprises a public corpus of news (with and without a hyper-partisan character), drawn from different websites between the years 2016 and 2018; the aim was to carry out a detailed analysis of the robustness of the algorithms under a qualitative analysis, from the human perspective of interpretation. For this qualitative analysis, a series of tasks were carried out based on structured questionnaires applied to undergraduate students at the University of São Paulo. Experiments with both synthetic data and real-world data showed algorithms with binary restrictions performed better than the others. An analysis of words that best represent clusters of news showed algorithms' difficulties in precisely defining, in the semantic sense, such clusters. The algorithm proposed in this work (WC-FNMTF) was submitted to several tasks and presented promising results. The task with humans revealed the superiority of the NBVD algorithm, followed by the WC-FNMTF.

Keywords: Co-clustering. Matrix Factorization. Human Interpretation.

Lista de figuras

Lista de algoritmos

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

ARI	*Adjusted Rand Index*
BinOvNMTF	*Overlapped Binary Non-negative Matriz Tri-Factorization*
BoW	*Bag-of-Words*
BVD	*Block Value Decomposition*
CA	*Clustering Accuracy*
FNMTF	*Fast Non-negative Matrix Tri Factorization*
MI	*Mutual Information*
NBVD	*Non-negative Block Value Decomposition*
NMF	*Non-negative Matrix Factorization*
NMI	*Normalized Mutual Information*
NMTF	*Non-negative Matrix Tri-Factorization*
ONMTF	*Orthogonal Non-negative Matrix Tri-Factorization*
ONM3F	*Orthogonal Non-negative Matrix Tri-Factorization*
OvNMTF	*Overlapping Non-negative Matrix Tri-Factorization*
RI	*Rand Index*
TF	*Term Frequency*
TF-IDF	*Term Frequency - Inverse Document Frequency*
WC-FNMTF	*Word Co-occurrence regularized Fast Non-negative Matrix Tri-Factorization*
WC-NMTF	*Word Co-occurrence regularized Non-negative Matrix Tri-Factorization*

Sumário

1 Introdução

O volume de dados textuais disponíveis nas últimas décadas, principalmente de forma online, tem crescido exponencialmente e esse fenômeno tem atraído a atenção de muitos pesquisadores. Navegar, explorar e organizar grandes corpus de textos em bibliotecas virtuais digitais é um exemplo de uma atividade comum. O desafio de estruturar e extrair automaticamente informações de tais tipos de dados tem crescido na mesma proporção que o volume de dados o tem e representa um amplo campo de pesquisa.

A estruturação e organização de textos podem ser apoiadas pelos resultados obtidos com a execução da tarefa de mineração de dados chamada agrupamento. Pesquisas em mineração de textos têm sido realizadas sobre tarefas de agrupamento (HAN; KAMBER; PEI, 2011), sobretudo, na exploração de dados em alta esparsidade e dimensionalidade.

Segundo o estudo de Jain, Murty e Flynn (1999), a análise de agrupamento pode ser vista como uma tarefa de organização de padrões em grupos cujos dados são mais similares entre si, enquanto padrões organizados em grupos distintos são mais dissimilares entre si. Pode-se dizer que a ideia por trás de modelos tradicionais de agrupamento é maximizar a similaridade intragrupos e minimizar a similaridade intergrupos. Todavia, o processo de agrupamento aplica-se apenas sobre os dados (os elementos) de um determinado conjunto, mas, dependendo do problema, existe também o interesse em analisar as características que descrevem esses dados. Coagrupamento pode ser visto como uma técnica que supre esse interesse. O objetivo do coagrupamento é encontrar subconjuntos de dados e de atributos, considerando os próprios dados e seus atributos descritivos. De forma geral, o processo de coagrupamento é semelhante ao processo de agrupamento, entretanto, aplicado simultaneamente sobre as linhas e colunas de uma matriz de dados (HARTIGAN, 1972; LONG; ZHANG; YU, 2005).

Coagrupamento organiza a informação de maneira mais detalhada do que o agrupamento o faz, haja vista que o primeiro considera duas dimensões de informações, o que oferece maior flexibilidade na definição dos grupos. Além disso, o processo de coagrupamento resulta em grupos de dados mais precisos, pois executa a análise simultânea dos dados (linhas) e de seus atributos (colunas).

Essa maneira de formular análise descritiva de dados tem sido promissora em problemas reais caracterizados por padrões subjetivos de interpretação, como no caso de

análise de imagens e de dados textuais (LEE; SEUNG, 1999; LONG; ZHANG; YU, 2005; YOO; CHOI, 2010; WANG *et al.*, 2011; HUANG; XU; LV, 2018). Em um contexto de aplicação de mineração de textos, cujo objetivo seja encontrar textos similares a outros textos, algoritmos de coagrupamento podem ser oportunos na identificação de palavras polissêmicas, por exemplo.

Diferentes algoritmos de coagrupamento foram propostos na literatura com aplicação para textos: NBVD (LONG; ZHANG; YU, 2005), ONM3F (DING *et al.*, 2006), ONMTF (YOO; CHOI, 2010), FNMTF (WANG *et al.*, 2011), BinOvNMTF (BRUNIALTI *et al.*, 2017), WC-NMTF (SALAH; AILEM; NADIF, 2018), OvNMTF (FREITAS JR. *et al.*, 2020). Eles baseiam-se em fatoração de matrizes, que é uma técnica de análise de dados capaz de extrair conhecimento de um objeto a partir do estudo de suas partes (LEE; SEUNG, 1999). Essa técnica é adequada para análise de dados diádicos (YOO; CHOI, 2010) e como estratégia de redução de dimensionalidade, já que compacta a matriz de dados em outras matrizes (CASALINO *et al.*, 2018). A representação tradicionalmente utilizada para textos é baseada no modelo de espaço vetorial (YOO; CHOI, 2010; WANG; OGIHARA, 2015; BRUNIALTI *et al.*, 2017; SALAH; AILEM; NADIF, 2018). Nesse modelo, cada documento é representado por um vetor e cada palavra do documento pode representar uma dimensão no espaço vetorial, portanto, o conjunto é dito de alta dimensionalidade, que também pode implicar em alta esparsidade e de difícil manipulação.

Alguns estudos também apresentam algoritmos de agrupamento baseados em fatoração de matrizes com aplicabilidade para textos (HOFMANN, 1999; LEE; SEUNG, 1999; XU; LIU; GONG, 2003; ALLAB; LABIOD; NADIF, 2016; ALZAHRANI *et al.*, 2016; AILEM; AGHILES; NADIF, 2017; CASALINO *et al.*, 2018). Tais métodos até permitem a extração de informações dos grupos para, por exemplo, identificar polissemia em textos, mas requerem etapas adicionais de pós-processamento, já que o agrupamento unilateral não providencia naturalmente grupos de atributos (YOO; CHOI, 2010).

Apesar do exposto, os estudos supracitados concentraram-se na capacidade dos algoritmos no agrupamento de dados e não trataram o aspecto da interpretabilidade dos grupos de textos sob o ponto de vista de análise qualitativa (LONG; ZHANG; YU, 2005; DING *et al.*, 2006; BRUNIALTI *et al.*, 2017; WANG *et al.*, 2011) ou o fizeram como um objetivo secundário do trabalho (YOO; CHOI, 2010; SALAH; AILEM; NADIF, 2018; FREITAS JR. *et al.*, 2020). O presente trabalho propôs um estudo exploratório e comparativo entre algoritmos de agrupamento e coagrupamento baseados em fatoração

de matrizes, tanto sob a ótica de avaliação quantitativa por meio de índices de validação de grupos e erro de reconstrução quanto sob a ótica de análise qualitativa por meio da avaliação dos resultados dos algoritmos sob uma perspectiva de interpretação humana.

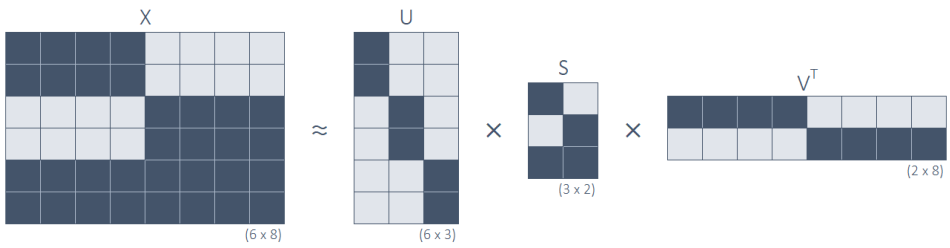
1.1 Definição do problema

Para definição do problema, faz-se necessário definir formalmente a tarefa de coagrupamento. Um conjunto de dados pode ser representado por uma matriz X de dimensão $n \times m$, em que $X \in \mathbb{R}^{n \times m}$, n representa o número de linhas da matriz e m representa o número de colunas. A matriz X compreende um conjunto de vetores de linhas $N = \{x_{1\cdot}, \ldots, x_{n\cdot}\}$ e um conjunto de vetores de colunas $M = \{x_{\cdot 1}, \ldots, x_{\cdot m}\}$, sendo x_{ij} um elemento dessa matriz. A tarefa de coagrupamento objetiva encontrar $k \times l$ cogrupos representados por submatrizes de X, sendo k o número de grupos de dados e l o número de grupos de atributos.

Uma maneira de visualizar essa representação matricial dos dados para coagrupamento é usar a ideia de dados diádicos. Segundo o estudo de Hofmann, Puzicha e Jordan (1999), dados diádicos referem-se a um domínio com dois conjuntos finitos de objetos $X = \{x_1, \ldots, x_i, \ldots, x_n\}$ e $Y = \{y_1, \ldots, y_j, \ldots, y_m\}$, cujas observações são na forma de díades (x_i, y_j). Uma díade pode consistir da coocorrência de x_i e y_j, por exemplo. No caso de textos, o conjunto X representa documentos e o conjunto Y representa palavras.

Fatorar uma matriz implica em encontrar duas ou mais matrizes que, ao serem agrupadas, recompõem a matriz original. Se três matrizes foram geradas durante o processo de fatoração, por exemplo, a matriz X poderia ser aproximada calculando USV^T, conforme ilustrado na figura 1. Mais detalhes sobre o processo de fatoração de matrizes serão fornecidos no decorrer deste texto.

Figura 1 – Ilustração do processo de fatoração tripla para aproximar a matriz original



Fonte: Waldyr Lourenço de Freitas Junior, 2023

As matrizes U, S e V proporcionam uma interpretação que favorece a exploração de dados textuais no processo de coagrupamento. A matriz U indica em qual dos k grupos foi alocado cada documento, a matriz V indica em qual dos l grupos foi alocada cada palavra e em especial a matriz S transmite uma noção de relação entre os grupos de linhas e os grupos de colunas por meio de um fator (peso), ou seja, a relação de um ou mais grupos de documentos com um, ou mais grupos de palavras.

O estudo de Yoo e Choi (2010) usou a matriz S para identificar palavras polissêmicas, entretanto, os autores não exploraram o uso da matriz S em outros algoritmos além do proposto por eles. O estudo de Salah, Ailem e Nadif (2018) explorou a preservação semântica das palavras introduzindo uma nova matriz no método, uma matriz de coocorrência de palavras, em uma fatoração adicional. Os autores apresentam uma análise das palavras mais representativas dos grupos de palavras, contudo, não exploraram o mesmo comportamento em outros métodos, para fins de comparação. O estudo de Freitas Jr. *et al.* (2020) também usou a matriz S objetivando relacionar tópicos e palavras em corpus de notícias e comparou os resultados do algoritmo proposto com o algoritmo ONMTF (YOO; CHOI, 2010).

Por fim, não foi encontrado um estudo amplo que explorasse as matrizes fatoradas, e comparasse os resultados dos diversos métodos de coagrupamento, sob o ponto de vista de interpretabilidade humana. Alguns estudos que exploraram diversos algoritmos, se limitaram a avaliá-los sob uma ótica quantitativa, para validação do agrupamento (SALAH; AILEM; NADIF, 2018; ABE; YADOHISA, 2019; LIU; HUA; CHEN, 2019; FEBRISSY *et al.*, 2022). Objetivando tratar essa lacuna, algumas questões de pesquisa foram avaliadas neste trabalho, detalhadas na seção 1.2.

1.2 Questões de pesquisa

O problema aqui estudado advém da natureza subjetiva da interpretação humana e da lacuna encontrada na literatura referente a essa tarefa no contexto de avaliação dos resultados de algoritmos de coagrupamento. Portanto, as questões a seguir foram avaliadas:

1. Considerando implementações baseadas em fatoração de matrizes, os algoritmos de coagrupamento podem produzir resultados de maior qualidade que os algoritmos de agrupamento, quando a avaliação da qualidade ocorre sob a perspectiva da interpretabilidade humana?

Como estratégia para responder a esta questão, uma questão derivada foi formulada:

   a) A capacidade dos algoritmos, medida de forma quantitativa, corresponde à capacidade deles quando passam por uma avaliação qualitativa, ou seja, os resultados bem avaliados segundo medidas quantitativas também são bem avaliados segundo as avaliações qualitativas?

2. No que diz respeito apenas aos diferentes algoritmos de coagrupamento baseados em fatoração de matrizes, há diferenças quanto à qualidade da interpretabilidade humana permitida pelos grupos que cada algoritmo produz?
3. Como extrair a informação contida nos grupos gerados pelos algoritmos e organizá-las para serem submetidas à avaliação qualitativa realizada por humanos?

1.3 Justificativa

Com relação ao problema proposto neste trabalho e suas respectivas questões de pesquisa, justifica-se que não foi encontrado na literatura um estudo que explorasse a interpretabilidade humana dos resultados dos algoritmos de coagrupamento. Como já exposto anteriormente, os estudos de Yoo e Choi (2010), Salah, Ailem e Nadif (2018) e Freitas Jr. *et al.* (2020) trataram essa questão de maneira secundária, dando uma ênfase maior na análise quantitativa dos algoritmos.

Sendo assim, um estudo exploratório de algoritmos de coagrupamento baseado em fatoração de matrizes sobre conjuntos de textos com diferentes características foi justificado como uma maneira de estudar o problema da interpretabilidade decorrente dos resultados de tais algoritmos.

1.4 Objetivos

O objetivo geral deste trabalho foi explorar sistematicamente um conjunto de algoritmos de coagrupamento baseados em fatoração de matrizes, com vista principalmente à interpretabilidade dos resultados produzidos por eles.

O processo para atingir o objetivo principal desse projeto permitiu delinear os objetivos específicos a seguir:

- Sistematizar formas de extração de informação de grupos dos resultados das fatorações dos diferentes algoritmos e organizar essas informações de modo que pessoas possam analisá-las.
- Propor um método adequado de análise qualitativa para submeter os resultados dos algoritmos à análise humana.

O cumprimento dos objetivos delineados propiciou o levantamento de limitações, prós e contras dos algoritmos avaliados, considerando as características de cada cenário.

1.5 Métodos

O trabalho aqui apresentado é de natureza aplicada, orientado a aprimorar soluções na área de coagrupamento de textos por meio de métodos baseados em fatoração de matrizes. Ele tem caráter explicativo, que segundo o estudo de Gil (2008), é caracterizado pelo objetivo de identificar fatores que colaboram para a ocorrência de fenômenos. Também pode ser considerado de gênero empírico, uma vez que os procedimentos metodológicos adotados foram a análise qualitativa com apoio de humanos e os experimentos computacionais. A abordagem metodológica que foi utilizada neste trabalho é a mista (quali-quanti), tanto pelo enfoque dado na questão de interpretabilidade humana dos resultados quanto pela questão da aferição da qualidade dos agrupamentos fornecidos pelos algoritmos, obtida via cálculo de índices de validação (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002).

As etapas abaixo foram definidas e executadas durante o projeto para atingir os objetivos deste trabalho, propostos da seção 1.4:

1. **Estudo da literatura**. Foi realizado um levantamento bibliográfico de estudos que pudessem apoiar o entendimento dos conceitos fundamentais e a construção do capítulo 2. Também foi realizada uma revisão exploratória, apoiada por um protocolo de busca e extração de dados, com o objetivo de levantar o estado da arte em métodos de agrupamento e coagrupamento baseados em fatoração de matrizes que aplicaram esses métodos sob uma ótica de interpretabilidade humana, visando consolidar a lacuna de pesquisa. O resultado dessa revisão está apresentado no capítulo 3.
2. **Escolha e implementação dos métodos utilizados**. Foram escolhidos os métodos de coagrupamento mais comuns encontrados na literatura, e variações deles, e um método mais atual que faz uso de princípios de coocorrência de palavras como recurso

adicional (SALAH; AILEM; NADIF, 2018). Todos eles são baseados em fatoração de matrizes. Os algoritmos que foram utilizados nos experimentos foram implementados pelo próprio autor deste trabalho de pesquisa em linguagem Python 3.0[1]. O método de agrupamento escolhido para comparação foi o *k-means*, implementado sobre a regra de fatoração de matrizes. Os métodos de coagrupamento utilizados foram: NBVD (LONG; ZHANG; YU, 2005), ONM3F (DING *et al.*, 2006), ONMTF (YOO; CHOI, 2010), FNMTF (WANG *et al.*, 2011), BinOvNMTF (BRUNIALTI *et al.*, 2017), WC-NMTF (SALAH; AILEM; NADIF, 2018), OvNMTF (FREITAS JR. *et al.*, 2020) e WC-FNMTF (método proposto neste trabalho). Cada um dos algoritmos está detalhado no capítulo 2.

3. **Análise da capacidade de agrupamento dos algoritmos**. Quanto maior for a capacidade de um algoritmo em agrupar dados, maior será a confiança na interpretabilidade decorrente dos resultados desse algoritmo. Essa capacidade foi analisada tanto com dados controlados quanto com dados do mundo real, com medidas interna e externa de validação de agrupamento. Algumas sub etapas tornaram-se necessárias para realizar essa validação: i) criação de conjuntos controlados de dados sintéticos, rotulados e com diferentes estruturas de cogrupos (MADEIRA; OLIVEIRA, 2004), ii) escolha e pré processamento de corpus com dados do mundo real, iii) testes dos diferentes algoritmos para os conjuntos previamente criados e para o conjunto público pré processado, variando seus parâmetros de entrada e iv) análise dos resultados dos algoritmos tanto do ponto de vista de agrupamento de linhas e agrupamento de colunas quanto de capacidade de reconstrução do conjunto de dados. A capacidade de agrupamento foi medida por meio do Índice de Rand Ajustado (HUBERT; ARABIE, 1985), do Índice *Silhouette* (ROUSSEEUW, 1987) e do erro de reconstrução. Essas medidas são comumente utilizadas na literatura (DING *et al.*, 2006; YOO; CHOI, 2010; WANG *et al.*, 2011; BRUNIALTI *et al.*, 2017; SALAH; AILEM; NADIF, 2018) para validar a capacidade de agrupamento de um algoritmo e estão mais bem detalhadas no capítulo 2.
4. **Análise da robustez dos algoritmos sob a ótica de interpretabilidade humana**. Interpretação de textos é um assunto subjetivo e traz um certo grau de complexidade para a realização automática dessa tarefa. Alguns estudos (YOO; CHOI, 2010; SALAH; AILEM; NADIF, 2018; FREITAS JR. *et al.*, 2020) sugerem

[1] https://www.python.org/download/releases/3.0/

abordagens para realizar extração de informações dos cogrupos com o objetivo de análise qualitativa. Este trabalho utilizou algumas destas abordagens: i) estudo dos vetores protótipos resultantes do processo de agrupamento, ii) avaliação das matrizes fatoradas, posto que para o processo de fatoração de três fatores uma das matrizes fatoradas possui a relação dos grupos de linhas com os grupos de colunas e iii) visualização e análise de representações textuais dos grupos e cogrupos. Foi utilizada a estratégia de questionário estruturado (SEIDMAN, 2006; GIBBS, 2009; LEITÃO; PRATES, 2017) para apoiar esta avaliação. Foi escolhido um corpus público de notícias (com e sem caráter de hiperpartidarismo) para análise da robustez dos algoritmos. Este corpus é uma coletânea de notícias extraídas de diferentes portais de notícias, majoritariamente entre os anos de 2016 e 2018. Os dados foram publicados como parte de uma tarefa do SemEval (*International Workshop on Semantic Evaluation*)[2] de 2019 e podem ser acessados pela plataforma Zenodo[3].

1.6 Organização do documento

Este documento é composto por cinco capítulos, incluindo a Introdução. No capítulo 2 são apresentados os principais conceitos e definições necessários referente à fatoração de matrizes, agrupamento, coagrupamento e medidas que foram utilizadas para análise quantitativa. O objetivo desse capítulo é fornecer um entendimento teórico dos assuntos que foram tratados nos demais capítulos. Uma discussão da literatura relacionada ao problema apresentado neste trabalho é feita no capítulo 3, com o objetivo de apresentar o estado da arte em coagrupamento. No capítulo 4 são apresentados os resultados dos experimentos. Esses experimentos ilustram a aplicabilidade dos métodos de coagrupamento. Por fim, o capítulo 5 apresenta as conclusões deste trabalho, as questões em aberto, as contribuições e as limitações.

[2] https://semeval.github.io/
[3] https://zenodo.org/record/1489920#.Y06AsnbMI2z

2 Conceitos fundamentais

Este capítulo apresenta os principais conceitos necessários para o entendimento deste trabalho de pesquisa. O presente capítulo foi dividido em quatro seções para introduzir os conceitos e notações referentes à fatoração de matrizes não negativas (seção 2.1), agrupamento (seção 2.2), coagrupamento (seção 2.3) e medidas de validação de agrupamento (seção 2.4). Os conceitos aqui apresentados podem ser suportados pelas propriedades matemáticas listadas no apêndice C.

2.1 Fatoração de matrizes não negativas

Fatoração de Matrizes Não Negativas, do inglês *Non-negative Matrix Factorization* (NMF), tem sido estudada como um método promissor de análise de dados a fim de extrair conhecimento sobre um item por meio da análise de suas partes (LEE; SEUNG, 1999; CASALINO *et al.*, 2018; IBRAHIM *et al.*, 2018). NMF foi proposta como uma alternativa a métodos mais tradicionais, como Análise de Componentes Principais e Quantização Vetorial, sobretudo pela sua distinção de restrições de não negatividade (LEE; SEUNG, 1999), que permite melhorar a interpretabilidade das informações extraídas (CHOO *et al.*, 2013; CASALINO; Del Buono; MENCAR, 2014).

Além disso, a forma de representar os dados tem se mostrado adequada para diversos contextos de agrupamento, como é o caso de dados diádicos. Dados diádicos referem-se a um domínio com dois conjuntos finitos de objetos, nos quais as observações são realizadas por meio de diádes, ou seja, pares de elementos de cada conjunto. Um exemplo muito comum de dados diádicos é a representação de textos em uma matriz de documentos por palavras, em que a relação delas é dada pela ocorrência de uma determinada palavra em um documento (LONG; ZHANG; YU, 2005).

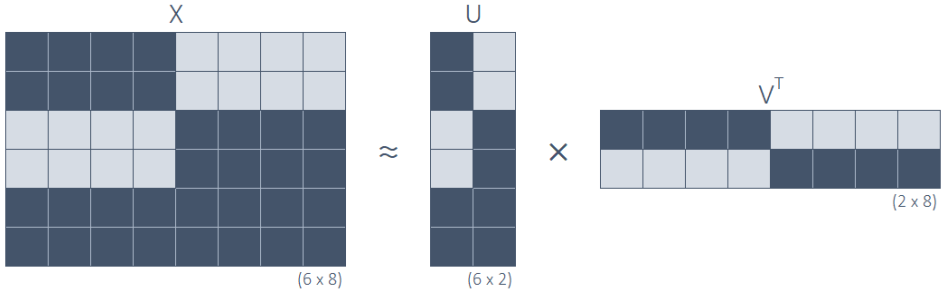
Algoritmos baseados em NMF têm como entrada uma matriz de dados $X \in \mathbb{R}_{+}^{n \times m}$, com n linhas que constituem um conjunto dos vetores de linhas $N = \{x_{1\cdot}, \ldots, x_{n\cdot}\}$, e m colunas que constituem um conjunto dos vetores de colunas $M = \{x_{\cdot 1}, \ldots, x_{\cdot m}\}$. A relação entre cada linha $x_{i\cdot}$ e cada coluna $x_{\cdot j}$ é representada por x_{ij}, com $i \in \{1, \ldots, n\}$ e $j \in \{1, \ldots, m\}$ (LEE; SEUNG, 1999). NMF é definida como a decomposição de uma

matriz não negativa em dois fatores, conforme apresentado na equação 1 (problema $\mathcal{F}_1$) e ilustrado na figura 2:

$$\begin{gathered}\mathcal{F}_1(U,V) = \min_{U,V} \|X - UV^T\|_F^2 \\ \text{sujeito a: } U \geq 0; V \geq 0,\end{gathered} \tag{1}$$

em que, $U \in \mathbb{R}_+^{n\times k}$, $V \in \mathbb{R}_+^{m\times k}$, $\|\cdot\|_F$ é a norma de *Frobenius*[1].

Figura 2 – Ilustração para o problema NMF considerando $n = 6$, $m = 8$ e $k = 2$. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

De acordo com o estudo de Yoo e Choi (2010), as colunas da matriz V correspondem aos vetores base para reconstrução da matriz de dados original, enquanto cada linha da matriz U representa uma codificação que determina com que extensão cada vetor base será utilizado no processo de reconstrução. Desta forma, as colunas da matriz V podem ser vistas como vetores protótipos para linhas da matriz de dados original.

2.2 Agrupamento

Segundo o estudo de Jain, Murty e Flynn (1999), a análise de agrupamento pode ser vista como uma tarefa de organização de grupos cujos elementos são mais similares entre si do que o são em relação aos elementos de outros grupos. De forma inversa, elementos organizados em grupos distintos são mais dissimilares entre si. Formalmente, dado um conjunto de dados representado pela matriz $X \in \mathbb{R}^{n\times m}$, de forma que os dados estão representados nas linhas e os seus atributos estão representados nas colunas, nos

[1] Uma das normas mais comuns em análise de dados é a norma Euclidiana, que no contexto de matrizes é também conhecida como norma *Frobenius*. Ela pode ser definida como $\|X\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$.

problemas de agrupamento, espera-se encontrar k grupos das N linhas de X, denotadas pelos subconjuntos $\mathcal{K}_i \subseteq N$, sendo $i \in \{1, \ldots, k\}$. $\mathcal{K} = \{K_1, \ldots K_k\}$ pode ser visto como os grupos de linhas resultantes que solucionam o problema de agrupamento.

2.2.1 *K-means*

O algoritmo de agrupamento *k-means* é um dos mais estudados na área de agrupamento. Este algoritmo objetiva encontrar k grupos que podem ser representados como um conjunto $\mathcal{K}' = \{\overrightarrow{K}_1, \ldots, \overrightarrow{K}_k\}$ de vetores protótipos. Cada um desses vetores protótipos está associado a um grupo do conjunto de dados e quantizam o espaço vetorial com relação ao erro de quantização mínimo.

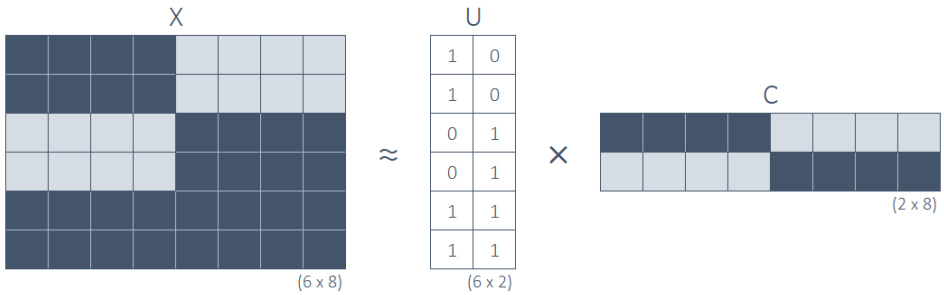
No contexto deste trabalho, assim como no trabalho de Ding e He (2005), o problema de agrupamento *k-means* é elaborado como a fatoração da matriz X em duas outras matrizes: U como uma matriz indicadora de grupos e C como uma matriz de vetores protótipos, tal que $X \approx UC$ e $\|X - UC\|_F^2$ define o erro de reconstrução da matriz original de dados. Esse problema é definido conforme $\mathcal{F}_2$, apresentado na equação 2:

$$\begin{aligned} \mathcal{F}_2(U, C) = \min_{U,C} \sum_{i=1}^{n} \sum_{p=1}^{k} u_{ip} \|x_{i\cdot} - c_{p\cdot}\|^2 = \min_{U,C} \|X - UC\|_F^2 \\ \text{sujeito a: } U \in \Psi^{n \times k};\ C \in \mathbb{R}^{k \times m};\ \textstyle\sum_{p=1}^{k} u_{ip} = 1 \ \ \forall i, \end{aligned} \tag{2}$$

em que $\Psi = \{0, 1\}$ e $\|\cdot\|_F$ é a norma de *Frobenius* para matrizes. O problema $\mathcal{F}_2$ pode ser visualizado graficamente na figura 3. A ilustração do problema $\mathcal{F}_2$ é muito próxima da ilustração do problema $\mathcal{F}_1$. Pode-se dizer que, da maneira como foi formulado, *k-means* é um caso especial de NMF, mas com algumas diferenças. No NMF, a matriz U aceita valores reais ao passo que no *k-means* a matriz aceita somente valores binários. Outro ponto é que o problema *k-means* não possui restrições de não negatividade na matriz C.

O algoritmo 1 é uma variação do algoritmo *k-means* tradicional, em que a atualização de C é derivada de forma matricial, criando um algoritmo iterativo com fatoração de matrizes.

Figura 3 – Ilustração do *k-means* por meio de fatoração de matrizes. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz. A matriz U é binária, conforme definição do problema.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

Algoritmo 1 Algoritmo *K-means* - Algoritmo baseado em fatoração de matrizes

function K-MEANS(X, k, t^{max})
  Initialize: $C^{(0)} \leftarrow \mathcal{U}(0,1)$, $U^{(0)} \leftarrow 0,1$ e $t \leftarrow 0$.
  while $(t \leq t^{max})$ e (não convergiu) **do**

$$C^{(t+1)} \leftarrow (U^{(t)^T} U^{(t)})^{-1} U^{(t)^T} X$$

$$(U^{(t+1)})_{ip} \leftarrow \begin{cases} 1 & p = \arg\min_{p' \in \{1,\ldots,k\}} \|\mathbf{x}_{i\cdot} - \mathbf{c}_{p'\cdot}^{(t+1)}\|^2 \\ 0 & \textit{caso contrário} \end{cases} \quad \forall i, p$$

    $t \leftarrow t + 1$
  end while
  return $U^{(t)}, C^{(t)}$
end function

Fonte: Ding e He (2005) e Brunialti (2016)

2.3 Coagrupamento

Coagrupamento pode ser visto como uma técnica de agrupamento de dados, semelhante ao processo de agrupamento, entretanto, aplicado simultaneamente sobre as linhas e colunas de uma matriz de dados (HARTIGAN, 1972). Formalmente, considere um conjunto de dados representado pela matriz $X \in \mathbb{R}^{n \times m}$. A matriz X compreende um conjunto de vetores de linhas $N = \{x_{1\cdot}, \ldots, x_{n\cdot}\}$ e um conjunto de vetores de colunas $M = \{x_{\cdot 1}, \ldots, x_{\cdot m}\}$. O objetivo é encontrar $k \times l$ cogrupos representados por submatrizes de X, denotados por $X_{K_p L_q}$, sendo k subconjuntos $K_p \subseteq N$, l subconjuntos $L_q \subseteq M$, $p \in \{1, \ldots, k\}$ e $q \in \{1, \ldots, l\}$. Pode-se dizer que nos problemas de coagrupamento, um cogrupo $X_{K_p L_q}$ é formado por um grupo de dados K_p e seus atributos L_q.

2.3.1 NBVD

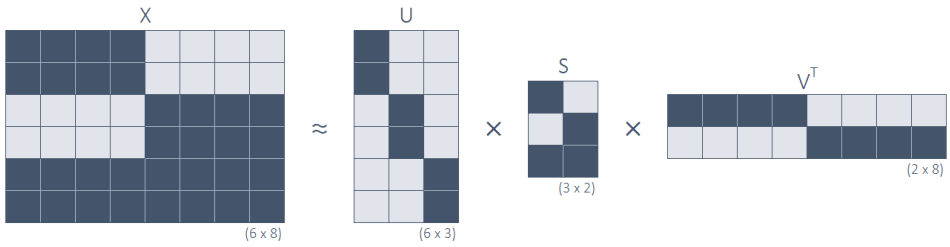
Decomposição de Valores em Blocos, do inglês *Block Value Decomposition* (BVD), busca por estruturas em blocos em uma matriz de dados e pode ser utilizada para análise de dados diádicos (LONG; ZHANG; YU, 2005). É uma técnica útil para soluções de coagrupamento uma vez que considera ambas as dimensões das matrizes de dados (linhas e colunas) simultaneamente. Esse mecanismo é realizado por meio da decomposição da matriz $X \in \mathbb{R}^{n \times m}$ em três outras matrizes (Problema $\mathcal{F}_3$): U como uma matriz de coeficientes de linhas, S como uma matriz de estrutura em blocos e V como uma matriz de coeficientes de colunas, conforme equação 3:

$$\begin{aligned} \mathcal{F}_3(U,S,V) &= \min_{U,S,V} \|X - USV^T\|_F^2 \\ \text{sujeito a: } & U \geq 0;\ V \geq 0, \end{aligned} \tag{3}$$

em que $U \in \mathbb{R}_+^{n \times k}$, $S \in \mathbb{R}^{k \times l}$ e $V \in \mathbb{R}_+^{m \times l}$.

O processo de fatoração de matrizes que resolve o problema $\mathcal{F}_3$ também pode ser ilustrado graficamente. A figura 4 ilustra esse processo.

Figura 4 – Ilustração do problema NBVD por meio de fatoração de matrizes. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

O problema $\mathcal{F}_3$ também é conhecido como NMTF, do inglês *Non-negative Matrix Tri-Factorization*, e é uma alternativa ao tradicional problema NMF, pois utiliza fatoração tripla de matrizes e naturalmente produz uma estrutura de coagrupamento. O trabalho de Long, Zhang e Yu (2005) propõe a seguinte interpretação para as matrizes resultantes da fatoração: S é uma representação compactada da matriz original X, a matriz US contém os vetores base para os grupos de colunas em X, a matriz SV^T contém os vetores base

para os grupos de linhas em X e as matrizes U e V denotam a associação de uma linha ou coluna ao seu respectivo grupo de linhas, ou colunas.

Desta forma, vetores protótipos podem ser extraídos para ambos os grupos, de linhas e de colunas, e a noção de coagrupamento pode ser explorada por meio da informação contida na matriz S, como apresentado no trabalho de Yoo e Choi (2010). O problema BVD restrito a uma matriz de dados exclusivamente positiva, isto é, $X \in \mathbb{R}_{+}^{n \times m}$, resulta na Decomposição de Valores em Blocos Não Negativos, do inglês *Non-negative Block Value Decomposition* (NBVD), proposto no trabalho de Long, Zhang e Yu (2005). O problema NBVD também foi explorado no presente trabalho e a implementação do processo de minimização dele está detalhada no algoritmo 2.

Algoritmo 2 Algoritmo NBVD - Decomposição de Valores em Blocos Não Negativos

function NBVD(X, k, l, itr^{max})
  Initialize: $U^{(0)} \leftarrow \mathcal{U}(0,1), V^{(0)} \leftarrow \mathcal{U}(0,1), S^{(0)} \leftarrow \mathcal{U}(0,1)$ e $t \leftarrow 0$.
  while $(t \leq itr^{max})$ e (não convergiu) **do**

$$U^{(t+1)} \leftarrow U^{(t)} \odot \frac{XV^{(t)}S^{(t)^T}}{U^{(t)}S^{(t)}V^{(t)^T}V^{(t)}S^{(t)^T}}$$

$$V^{(t+1)} \leftarrow V^{(t)} \odot \frac{X^TU^{(t+1)}S^{(t)}}{V^{(t)}S^{(t)^T}U^{(t+1)^T}U^{(t+1)}S^{(t)}}$$

$$S^{(t+1)} \leftarrow S^{(t)} \odot \frac{U^{(t+1)^T}XV^{(t+1)}}{U^{(t+1)^T}U^{(t+1)}S^{(t)}V^{(t+1)^T}V^{(t+1)}}$$

    $t \leftarrow t + 1$
  end while
  return $U^{(t)}, S^{(t)}, V^{(t)}$
end function

Fonte: Long, Zhang e Yu (2005) e Brunialti (2016)

2.3.2 ONM3F e ONMTF

O problema $\mathcal{F}_4$, apresentado na equação 4, foi proposto inicialmente no trabalho de Ding *et al.* (2006) e é chamado de Fatoração Ortogonal Tripla de Matrizes Não Negativas, do inglês *Orthogonal Non-negative Matrix Tri-Factorization* (ONMTF). Nesse problema, além das restrições de não negatividade e fatoração tripla já discutidas nos problemas anteriores, foram adicionadas restrições de ortogonalidade nas matrizes U e V, respectivamente: $U^TU = I$ e $V^TV = I$, em que I é a matriz identidade. Essas restrições

limitam o problema de fatoração de $X \approx USV^T$ para um número bem menor de soluções possíveis:

$$\begin{aligned} \mathcal{F}_4(U,S,V) &= \min_{U,S,V} \|X - USV^T\|_F^2 \\ \text{sujeito a: } U \geq 0;\ S \geq 0;\ V \geq 0;\ U^TU = I;\ V^TV = I, \end{aligned} \tag{4}$$

em que $U \in \mathbb{R}_+^{n\times k}, S \in \mathbb{R}_+^{k\times l}$ e $V \in \mathbb{R}_+^{m\times l}$, $\|\cdot\|_F$ é a norma *Frobenius* para matrizes e $\|X - USV^T\|_F^2$ fornece o erro de reconstrução.

O trabalho de Yoo e Choi (2010) também explorou o problema ONMTF e apresentou um novo método baseado no cálculo do gradiente. As regras de atualizações multiplicativas propostas se basearam em uma superfície que preserva as restrições de ortogonalidade, conhecida como Variedade de Stiefel[2] (*Stiefel Manifold*). A nomenclatura que está sendo utilizada neste trabalho é a mesma que foi utilizada no trabalho de Yoo e Choi (2010): ONM3F para o algoritmo proposto no trabalho de Ding *et al.* (2006) e ONMTF para o próprio algoritmo dos autores, ambos baseados no mesmo problema ($\mathcal{F}_4$). O algoritmo 3 detalha o ONM3F e o algoritmo 4 detalha o ONMTF.

Graficamente, o problema $\mathcal{F}_4$ pode ser ilustrado conforme a figura 4, figura em que foi ilustrado o NBVD. Uma representação gráfica para o problema $\mathcal{F}_4$ é igual à representação gráfica do NBVD porque as restrições de ortogonalidade não podem ser representadas nessa ilustração.

2.3.3 OvNMTF

O problema $\mathcal{F}_5$ foi introduzido no trabalho de Brunialti (2016) e apresentado para a comunidade científica no trabalho de Freitas Jr. *et al.* (2020). Intitulado como Fatoração Tripla de Matrizes Não Negativas Sobrepostas, do inglês *Overlapping Non-negative Matrix Tri-Factorization* (OvNMTF), é baseado nas premissas do NBVD, apresentado na equação 3 (problema $\mathcal{F}_3$), e é proposto com o objetivo de analisar os grupos de atributos (colunas) de forma independente dos grupos de dados (linhas), assumindo a existência de um número k de matrizes V, em que k é o número de grupos de dados. A existência dessas k matrizes permite maior flexibilidade para encontrar os grupos de dados, pois cria uma relação única

[2] O cálculo do gradiente para as atualizações multiplicativas é feito sobre uma superfície com restrições que preserva a ortogonalidade. Essa superfície é conhecida como *Variedade de Stiefel* - um conjunto de matrizes ortonormais.

Algoritmo 3 Algoritmo ONM3F - Fatoração Ortogonal Tripla de Matrizes Não Negativas

function ONM3F(X, k, l, t^{max})
  Initialize: $U^{(0)} \leftarrow \mathcal{U}(0,1), V^{(0)} \leftarrow \mathcal{U}(0,1), S^{(0)} \leftarrow \mathcal{U}(0,1)$ e $t \leftarrow 0$.
  while $(t \leq t^{max})$ e (não convergiu) **do**

$$U^{(t+1)} \leftarrow U^{(t)} \odot \sqrt{\frac{XV^{(t)}S^{(t)^T}}{U^{(t)}U^{(t)^T}XV^{(t)}S^{(t)^T}}}$$

$$V^{(t+1)} \leftarrow V^{(t)} \odot \sqrt{\frac{X^TU^{(t+1)}S}{V^{(t)}V^{(t)^T}X^TU^{(t+1)}S^{(t)}}}$$

$$S^{(t+1)} \leftarrow S^{(t)} \odot \sqrt{\frac{U^{(t+1)^T}XV^{(t+1)}}{U^{(t+1)^T}U^{(t+1)}S^{(t)}V^{(t+1)^T}V^{(t+1)}}}$$

    $t \leftarrow t + 1$
  end while
  return $U^{(t)}, S^{(t)}, V^{(t)}$
end function

Fonte: Ding *et al.* (2006) e Brunialti (2016)

Algoritmo 4 Algoritmo ONMTF - Fatoração Ortogonal Tripla de Matrizes Não Negativas baseado na teoria de derivação na superfície com restrições (Variedade Stiefel)

function ONMTF(X, k, l, t^{max})
  Initialize: $U^{(0)} \leftarrow \mathcal{U}(0,1), V^{(0)} \leftarrow \mathcal{U}(0,1), S^{(0)} \leftarrow \mathcal{U}(0,1)$ e $t \leftarrow 0$.
  while $(t \leq t^{max})$ e (não convergiu) **do**

$$U^{(t+1)} \leftarrow U^{(t)} \odot \frac{XV^{(t)}S^{(t)^T}}{U^{(t)}S^{(t)}V^{(t)^T}X^TU^{(t)}}$$

$$V^{(t+1)} \leftarrow V^{(t)} \odot \frac{X^TU^{(t+1)}S^{(t)}}{V^{(t)}S^{(t)^T}U^{(t+1)^T}XV^{(t)}}$$

$$S^{(t+1)} \leftarrow S^{(t)} \odot \frac{U^{(t+1)^T}XV^{(t+1)}}{U^{(t+1)^T}U^{(t+1)}S^{(t)}V^{(t+1)^T}V^{(t+1)}}$$

    $t \leftarrow t + 1$
  end while

$$U^{(t)} \leftarrow U^{(t)}\,\mathrm{diag}(S^{(t)}\,\mathrm{diag}(\mathbf{1}^TV^{(t)})\mathbf{1})$$

$$V^{(t)} \leftarrow V^{(t)}\,\mathrm{diag}(\mathbf{1}^T\,\mathrm{diag}(\mathbf{1}^TU^{(t)})S^{(t)})$$

  return $U^{(t)}, S^{(t)}, V^{(t)}$
end function

Fonte: Yoo e Choi (2010) e Brunialti (2016)

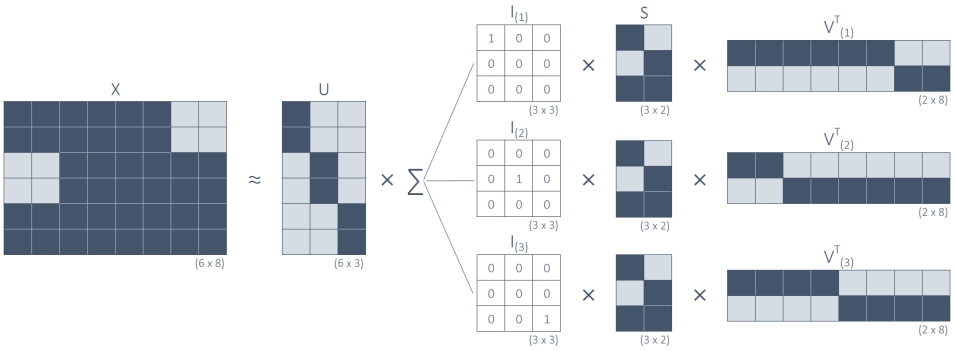
de cada matriz V a um único grupo de dados (A ideia é que cada grupo de dados seja otimizado simultaneamente à otimização de um grupo de atributos exclusivo para ele). Essa relação é possível por meio de matrizes $I_{(p)}$ introduzidas pelos autores no processo de fatoração, intituladas matrizes seletoras. A matriz $I_{(p)}$ contém zeros em todos os seus elementos, salvo no $i_{(p)_{pp}}$ elemento, que é igual a um, e proporciona o efeito da relação supracitada. O problema OvNMTF é formulado conforme equação 5:

$$\begin{aligned} \mathcal{F}_5(U, S, V_{(1)}, \ldots, V_{(k)}) = \min_{U,S,V_{(1)},\ldots,V_{(k)}} \|X - U \sum_{p=1}^{k} I_{(p)} S V_{(p)}^T\|_F^2 \\ \text{sujeito a: } U \geq 0;\ S \geq 0;\ V_{(p)} \geq 0, \forall p, \end{aligned} \tag{5}$$

em que $U \in \mathbb{R}_+^{n \times k}$, $S \in \mathbb{R}_+^{k \times l}$, $V_{(p)} \in \mathbb{R}_+^{m \times l}$, $p \in \{1, \ldots, k\}$ como índice para o conjunto de matrizes $\{V_{(1)}, \ldots, V_{(k)}\}$, $I_{(p)} \in \{0, 1\}^{k \times k}$ é a matriz seletora tendo o elemento $i_{(p)_{pp}} = 1$ e os demais elementos iguais a zero. A matriz $\sum_{p=1}^{k} I_{(p)} S V_{(p)}^T$ contém os vetores base para os grupos de linhas da matriz X.

O algoritmo homônimo que soluciona o problema OvNMTF foi desenhado para naturalmente tratar sobreposição de colunas durante o processo de coagrupamento, entretanto, é um algoritmo complexo em termos de custo computacional devido ao maior número de matrizes envolvidas no processo de fatoração. A figura 5 representa uma ilustração do problema $\mathcal{F}_5$ e o algoritmo 5 detalha as regras de atualização deste problema.

Figura 5 – Ilustração do problema OvNMTF por meio de fatoração de matrizes. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz. As matrizes $I_{(p)}$ são binárias e possuem as peculiaridades já expostas.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

Algoritmo 5 Algoritmo OvNMTF - Fatoração Tripla de Matrizes Não Negativas Sobrepostas

function OvNMTF(X, k, l, t^{max})
  Initialize: $U^{(0)} \leftarrow \mathcal{U}(0,1), S^{(0)} \leftarrow \mathcal{U}(0,1), V_{(p)}^{(0)} \leftarrow \mathcal{U}(0,1), \forall p$ e $t \leftarrow 0$.
  while $(t \leq t^{max})$ e (não convergiu) **do**

$$U^{(t+1)} \leftarrow U^{(t)} \odot \frac{\sum_{p=1}^{k} X V_{(p)}^{(t)} S^{(t)^T} I_{(p)}}{\sum_{p=1}^{k} \sum_{p'=1}^{k} U^{(t)} I_{(p)} S^{(t)} V_{(p)}^{(t)^T} V_{(p')}^{(t)} S^{(t)^T} I_{(p')}}$$

    for $p \leftarrow 1$ até k **do**

$$V_{(p)}^{(t+1)} \leftarrow V_{(p)}^{(t)} \odot \frac{X^T U^{(t+1)} I_{(p)} S^{(t)}}{\sum_{p'=1}^{k} V_{(p')} S^T I_{(p')} U^T U I_{(p)} S}$$

    end for

$$S^{(t+1)} \leftarrow S^{(t)} \odot \frac{\sum_{p=1}^{k} I_{(p)} U^{(t+1)^T} X V_{(p)}^{(t+1)}}{\sum_{p=1}^{k} \sum_{p'=1}^{k} I_{(p)} U^{(t+1)^T} U^{(t+1)} I_{(p')} S^{(t)} V_{(p')}^{(t+1)^T} V_{(p)}^{(t+1)}}$$

    $t \leftarrow t + 1$
  end while
  return $U^{(t)}, S^{(t)}, V_{(1)}^{(t)}, \ldots, V_{(k)}^{(t)}$
end function

Fonte: Brunialti (2016)

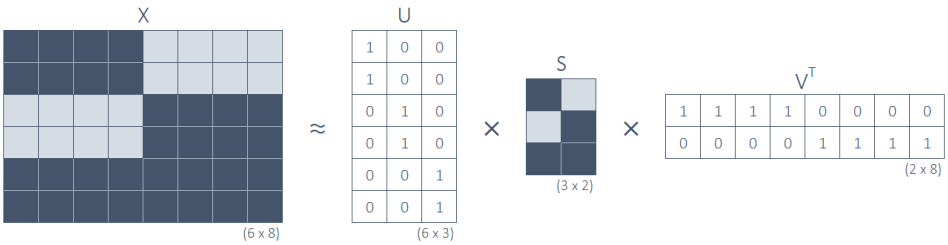
2.3.4 FNMTF

O problema $\mathcal{F}_6$, formulado na equação 6 e ilustrado na figura 6, foi proposto no trabalho de Wang *et al.* (2011) como uma alternativa para tratar a questão de desempenho dos métodos baseados em fatoração de matrizes não negativas. Tais métodos realizam multiplicações matriciais massivamente e o custo computacional associado a eles é alto. Intitulado pelos autores de Fatoração Tripla Rápida de Matrizes Não Negativas, do inglês *Fast Non-negative Matrix Tri Factorization* (FNMTF), a abordagem de FNMTF é fatorar uma matriz de dados X em três outras matrizes: U como uma matriz indicadora de grupos de linhas, S como uma matriz que relaciona os grupos de linhas e os grupos de colunas e V como uma matriz indicadora de grupos de colunas:

$$\begin{gathered} \mathcal{F}_6(U,S,V) = \min_{U,S,V} \|X - USV^T\|_F^2 \\ \text{sujeito a: } U \in \Psi^{n\times k};\ V \in \Psi^{m\times l} \\ \sum_{p=1}^{k} u_{ip} = 1, \forall i;\ \sum_{q=1}^{l} v_{jq} = 1, \forall j, \end{gathered} \tag{6}$$

em que $S \in \mathbb{R}_{+}^{k \times l}$, $\Psi = \{0, 1\}$ e $\| \cdot \|_F$ é a norma de *Frobenius*, para matrizes.

Figura 6 – Ilustração do problema FNMTF por meio de fatoração de matrizes. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz. As matrizes U e V são binárias, conforme definição do problema.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

Semelhante ao algoritmo *k-means* previamente apresentado, o algoritmo FNMTF possui regras iterativas de atualização, detalhadas no algoritmo 6.

Algoritmo 6 Algoritmo FNMTF - Fatoração Tripla Rápida de Matrizes Não Negativas

function FNMTF(X, k, l, t^{max})
  Initialize: $U^{(0)} \leftarrow 0, 1 \mid \sum_{p=1}^{k} u_{ip} = 1$, $V^{(0)} \leftarrow 0, 1 \mid \sum_{q=1}^{l} v_{jq} = 1, \forall i, j$, $S^{(0)} \leftarrow \mathcal{U}(0, 1)$ e $t \leftarrow 0$.
  while $(t \leq t^{max})$ e (não convergiu) **do**

$$S^{(t+1)} \leftarrow (U^{(t)^T} U^{(t)})^{-1} U^{(t)^T} X V^{(t)} (V^{(t)^T} V^{(t)})^{-1}$$

$$\widetilde{V} \leftarrow S^{(t+1)} V^{(t)^T}$$

$$(U^{(t+1)})_{ip} \leftarrow \begin{cases} 1 & p = \arg\min_{p' \in \{1,\ldots,k\}} \|\mathbf{x}_{i\cdot} - \widetilde{\mathbf{v}}_{p'\cdot}\|^2 \\ 0 & \textit{caso contrário} \end{cases} \quad \forall i, p$$

$$\widetilde{U} \leftarrow U^{(t+1)} S^{(t+1)}$$

$$(V^{(t+1)})_{jq} \leftarrow \begin{cases} 1 & q = \arg\min_{q' \in \{1,\ldots,l\}} \|\mathbf{x}_{\cdot j} - \widetilde{\mathbf{u}}_{\cdot q'}\|^2 \\ 0 & \textit{caso contrário} \end{cases} \quad \forall j, q$$

    $t \leftarrow t + 1$
  end while
  return $U^{(t)}, S^{(t)}, V^{(t)}$
end function

Fonte: Wang *et al.* (2011) e Brunialti (2016)

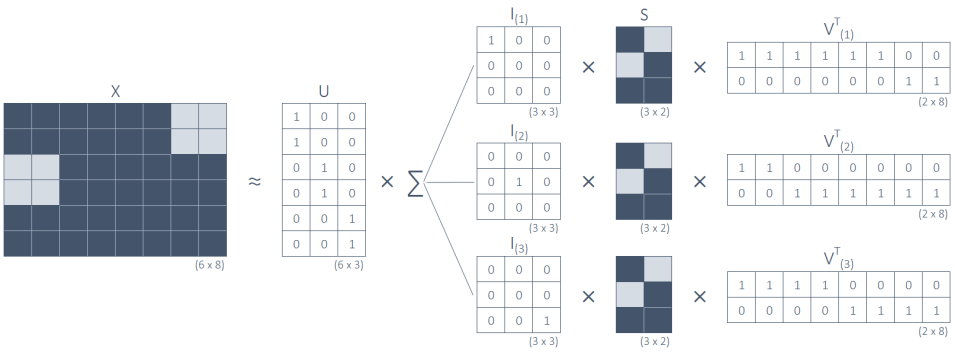
2.3.5 BinOvNMTF

O problema denominado Fatoração Binária Tripla de Matrizes Não Negativas com Sobreposição, do inglês *Overlapped Binary Non-negative Matriz Tri-Factorization* (BinOvNMTF), aqui definido como $\mathcal{F}_7$, foi proposto no trabalho de Brunialti (2016) e apresentado à comunidade científica no trabalho de Brunialti *et al.* (2017). Ele é baseado nas premissas do FNMTF, problema apresentado na equação 6 (problema $\mathcal{F}_6$). O detalhamento do problema está formulado na equação 7:

$$\begin{aligned} \mathcal{F}_7(U, S, V_{(1)}, \ldots, V_{(k)}) = \min_{U,S,V_{(1)},\ldots,V_{(k)}} \|X - U \sum_{p=1}^{k} I_{(p)} S V_{(p)}^T\|_F^2 \\ \text{sujeito a: } U \in \Psi^{n\times k};\ V_{(p)} \in \Psi^{m\times l}, \forall p; \\ \textstyle\sum_{p=1}^{k} u_{ip} = 1, \forall i;\ \sum_{q=1}^{l} v_{(p)_{jq}} = 1, \forall p, j, \end{aligned} \tag{7}$$

em que $S \in \mathbb{R}_+^{k\times l}$, $\Psi = \{0, 1\}$, $p \in \{1, \ldots, k\}$ e $q \in \{1, \ldots, l\}$ são índices que iteram sobre o número de grupos de linhas e colunas, respectivamente, $I_{(p)} \in \{0, 1\}^{k\times k}$ é a matriz seletora tendo $i_{(p)_{pp}} = 1$ e as demais entradas iguais a zero, e $\|\cdot\|_F$ é a norma de *Frobenius* para matrizes. A matriz seletora $I_{(p)}$ desempenha o mesmo papel que a matriz seletora introduzida na seção 2.3.3. A matriz $UI_{(p)}S$, $\forall p \in \{1, \ldots, k\}$ contém os vetores base para os grupos de atributos e a matriz $\sum_{p=1}^{k} I_{(p)} S V_{(p)}^T$ contém os vetores base para os grupos de dados. O problema $\mathcal{F}_7$ está ilustrado na figura 7.

Figura 7 – Ilustração do problema BinOvNMTF por meio de fatoração de matrizes. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz. As matrizes U e $V_{(p)}$ são binárias, conforme definição do problema. As matrizes $I_{(p)}$ também são binárias e possuem as peculiaridades já expostas.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

As regras iterativas de atualização para o algoritmo BinOvNMTF estão detalhadas no algoritmo 7, semelhante às que foram apresentadas para o algoritmo FNMTF.

Algoritmo 7 Algoritmo BinOvNMTF - Fatoração Binária Tripla de Matrizes Não Negativas com Sobreposição

function BinOvNMTF(X, k, l, t^{max})
  Initialize: $U^{(0)} \leftarrow 0,1 \mid \sum_{p=1}^{k} u_{ip} = 1$, $V_{(p)}^{(0)} \leftarrow 0,1 \mid \sum_{q=1}^{l} v_{(p)_{jq}} = 1, \forall i,j,p$ e $t \leftarrow 0$.
  while $(t \leq t^{max})$ e (não convergiu) **do**

$$S^{(t+1)} \leftarrow \sum_{p=1}^{k} I_{(p)} (U^{(t)^T} U^{(t)})^{-1} U^{(t)^T} X V_{(p)}^{(t)} (V_{(p)}^{(t)^T} V_{(p)}^{(t)})^{-1}$$

$$\widetilde{V} \leftarrow \sum_{p=1}^{k} I_{(p)} S^{(t+1)} V_{(p)}^{(t)^T}$$

$$(U^{(t+1)})_{ip} \leftarrow \begin{cases} 1 & p = \arg\min_{p' \in \{1,\ldots,k\}} \|\mathbf{x}_{i\cdot} - \widetilde{\mathbf{v}}_{p'\cdot}\|^2 \\ 0 & \textit{caso contrário} \end{cases} \quad \forall i,p$$

$$\widetilde{U}_{(p)} \leftarrow U^{(t+1)} I_{(p)} S^{(t+1)}, \forall p$$

$$(V_{(p)}^{(t+1)})_{jq} \leftarrow \begin{cases} 1 & q = \arg\min_{q' \in \{1,\ldots,l\}} \|\mathbf{x}_{\cdot j} - \widetilde{\mathbf{u}}_{(p)\cdot q'}\|^2 \\ 0 & \textit{caso contrário} \end{cases} \quad \forall j,p,q$$

    $t \leftarrow t+1$
  end while
  return $U^{(t)}, S^{(t)}, V_{(1)}^{(t)}, \ldots, V_{(k)}^{(t)}$
end function

Fonte: Brunialti (2016)

2.3.6 WC-NMTF

O problema $\mathcal{F}_8$ foi proposto no trabalho de Salah, Ailem e Nadif (2018). Ele é formulado na equação 8 e é intitulado como Fatoração Tripla de Matrizes Não Negativas Regularizada com Coocorrência de Palavras, do inglês *Word Co-occurrence regularized Non-negative Matrix Tri-Factorization* (WC-NMTF):

$$\begin{aligned} \mathcal{F}_8(U,S,V,Q) = \min_{U,S,V,Q} \frac{1}{2} \|X - USV^T\|_F^2 + \frac{\lambda}{2} \|M - VQ^T\|_F^2 \\ \text{sujeito a: } U \geq 0;\ S \geq 0;\ V \geq 0;\ M \geq 0;\ Q \geq 0, \end{aligned} \tag{8}$$

em que $U \in \mathbb{R}_+^{n\times k}, S \in \mathbb{R}_+^{k\times l}$, $V \in \mathbb{R}_+^{m\times l}$, $M \in \mathbb{R}_+^{m\times m}$ e $Q \in \mathbb{R}_+^{m\times l}$. A matriz Q desempenha um papel de fator extra, também chamado de fator de contexto, para decomposição de M, λ é um parâmetro de regularização e $\|\cdot\|_F$ é a norma *Frobenius* para matrizes.

A matriz M é a base do algoritmo WC-NMTF. Inicialmente, é definida uma matriz $C \in \mathbb{R}_+^{m\times m}$ que codifica o número de vezes que cada par de palavras aparece no mesmo contexto, ou seja, uma matriz de coocorrência. Dada a matriz C, a Informação Mútua Pontual (PMI) entre uma palavra w_j e outra palavra $w_{j'}$ pode ser estimada como:

$$PMI(w_j, w_{j'}) = \log \frac{c_{jj'} \times c_{..}}{c_{j.} \times c_{.j'}}, \tag{9}$$

em que $c_{jj'}$ é a coocorrência das palavras w_j e $w_{j'}$, $c_{..} = \sum_{p=1}^{j}\sum_{q=1}^{j'} c_{pq}, c_{j.} = \sum_{q=1}^{j'} c_{jq}$ e $c_{.j'} = \sum_{p=1}^{j} c_{pj'}$. Em seguida, com o argumento de que a matriz PMI é densa e de alta dimensionalidade para tratar o problema, os autores propõem uma nova transformação para uma matriz PMI Esparsa Deslocada Positiva, do inglês *Sparse Shifted Positive PMI* (SPPMI). Portanto, a matriz $M \in \mathbb{R}_+^{m\times m}$ é transformada conforme equação 10, em que $m_{jj'}$ representa os elementos da matriz:

$$m_{jj'} = \max\{PMI(w_j, w_{j'}) - \log(N), 0\}, \tag{10}$$

em que N é um hiperparâmetro que controla o nível de esparsidade de M (LEVY; GOLDBERG, 2014).

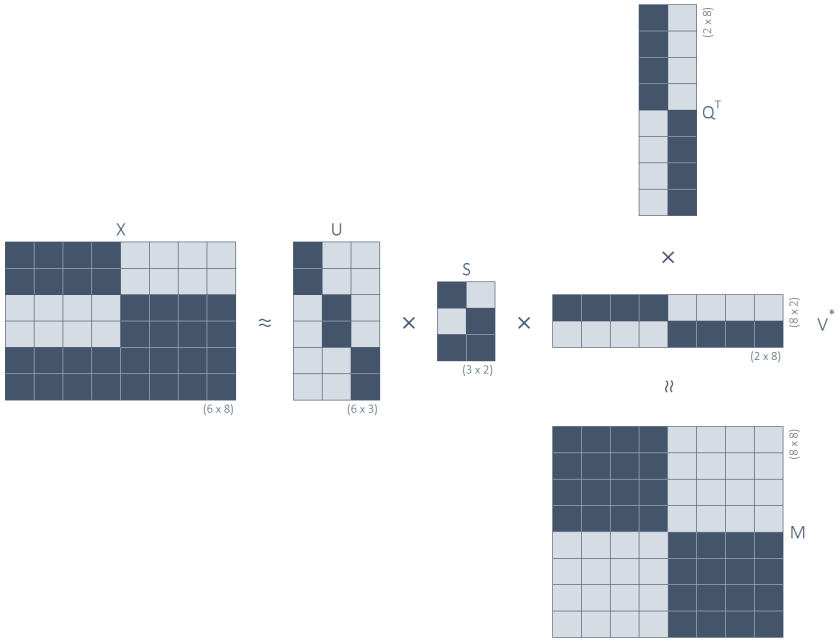
A figura 8 ilustra o problema $\mathcal{F}_8$, detalhado na equação 8. Para representar o método em um único gráfico, V^* possui a seguinte definição: $V^* = V^T$ para $X \approx USV^*$ e $V^* = V$ para $M \approx V^*Q^T$.

A ideia formulada nesse problema ($\mathcal{F}_8$) é maximizar a similaridade entre as palavras, mapeada aqui como coocorrência em um dado contexto, preservando esse relacionamento entre elas para melhorar a qualidade de um coagrupamento. O algoritmo 8 detalha o problema WC-NMTF.

2.3.7 WC-FNMTF

O problema $\mathcal{F}_{11}$ foi proposto neste trabalho. Ele é formulado na equação 11 e é intitulado como Fatoração Tripla Rápida de Matrizes Não Negativas Regularizada com

Figura 8 – Ilustração do problema WC-NMTF por meio de fatoração de matrizes. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz. $X \approx USV^*$, tendo $V^* = V^T$, e $M \approx V^*Q^T$, tendo $V^* = V$.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

Coocorrência de Palavras, do inglês *Word Co-occurrence regularized Fast Non-negative Matrix Tri-Factorization* (WC-FNMTF):

$$\begin{aligned} \mathcal{F}_{11}(U,S,V,Q) = \min_{U,S,V,Q} \frac{1}{2}\left\|X - USV^T\right\|_F^2 + \frac{\lambda}{2}\left\|M - VQ^T\right\|_F^2 \\ \text{sujeito a: } U \in \Psi^{n\times k};\ V \in \Psi^{m\times l};\ S \geq 0;\ M \geq 0;\ Q \geq 0 \\ \textstyle\sum_{p=1}^{k} u_{ip} = 1, \forall i;\ \sum_{q=1}^{l} v_{jq} = 1, \forall j, \end{aligned} \tag{11}$$

em que $S \in \mathbb{R}_+^{k\times l}$, $M \in \mathbb{R}_+^{m\times m}$, $Q \in \mathbb{R}_+^{m\times l}$, $\Psi = \{0,1\}$ e $\|\cdot\|_F$ é a norma de *Frobenius*, para matrizes. A matriz Q, a matriz M e o parâmetro λ possuem a mesma definição do problema definido na equação 8. As demais definições referentes à construção da matriz M, baseadas nas matrizes PMI e SPPMI, também seguem o que já foi detalhado na seção 2.3.6. A figura 9 ilustra o problema $\mathcal{F}_{11}$, apresentado na equação 11.

A motivação para este novo método é reunir as vantagens do FNMTF e do WC-NMTF em um único algoritmo, que é, respectivamente, diminuir a complexidade de tempo do algoritmo e maximizar a similaridade entre as palavras, preservando o relacionamento entre elas, e assim melhorar a qualidade de um coagrupamento. Não há restrições nas

Algoritmo 8 Algoritmo WC-NMTF - Fatoração Tripla de Matrizes Não Negativas Regularizada com Coocorrência de Palavras

function WC-NMTF(X, k, l, C, itr^{max})
  Initialize: $U^{(0)} \leftarrow \mathcal{U}(0,1), V^{(0)} \leftarrow \mathcal{U}(0,1), S^{(0)} \leftarrow \mathcal{U}(0,1), Q^{(0)} \leftarrow \mathcal{U}(0,1)$, $\lambda \leftarrow 1$, $M \leftarrow generateSPPMI(C)$ e $t \leftarrow 0$.
  while $(t \leq itr^{max})$ e (não convergiu) **do**

$$U^{(t+1)} \leftarrow U^{(t)} \odot \frac{XV^{(t)}S^{(t)^T}}{U^{(t)}S^{(t)}V^{(t)^T}V^{(t)}S^{(t)^T}}$$

$$V^{(t+1)} \leftarrow V^{(t)} \odot \frac{(X^TU^{(t+1)}S^{(t)} + \lambda MQ^{(t)})}{V^{(t)}(S^{(t)^T}U^{(t+1)^T}U^{(t+1)}S^{(t)} + \lambda Q^{(t)^T}Q^{(t)})}$$

$$S^{(t+1)} \leftarrow S^{(t)} \odot \frac{U^{(t+1)^T}XV^{(t+1)}}{U^{(t+1)^T}U^{(t+1)}S^{(t)}V^{(t+1)^T}V^{(t+1)}}$$

$$Q^{(t+1)} \leftarrow Q^{(t)} \odot \frac{M^TV^{(t+1)}}{Q^{(t)}V^{(t+1)^T}V^{(t+1)}}$$

    $t \leftarrow t + 1$
  end while
  return $U^{(t)}, S^{(t)}, V^{(t)}, Q^{(t)}$
end function

Fonte: Salah, Ailem e Nadif (2018)

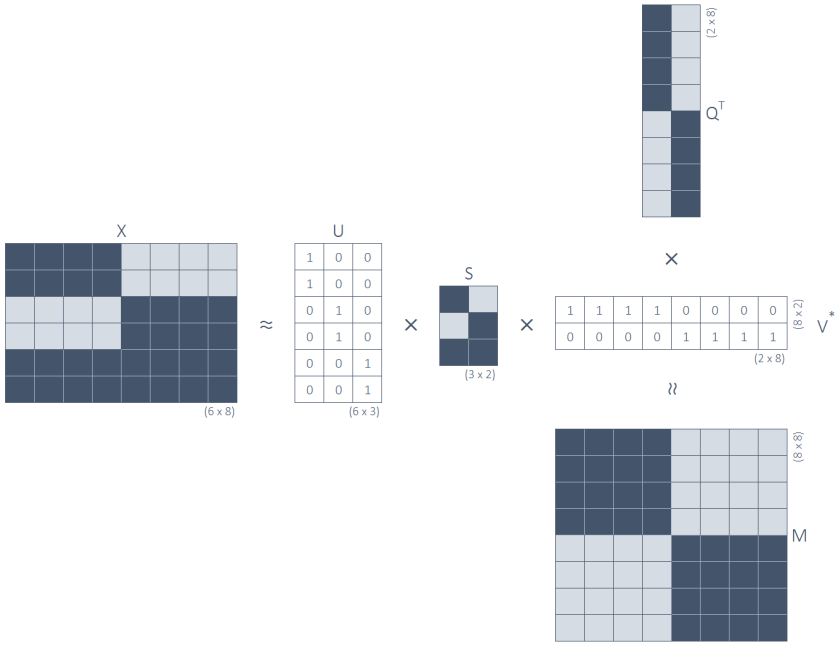
matrizes S e Q além das restrições de positividade que é assegurada pela positividade da matriz X. Desta maneira, é possível encontrar uma regra de atualização para S e para Q e consequentemente a minimização do problema $\mathcal{F}_{11}$:

$$\begin{aligned}
\nabla_S\mathcal{F}_{11} &= X - USV^T &&= 0\\
&\Longrightarrow USV^T &&= X\\
&\Longrightarrow U^TUSV^TV &&= U^TXV\\
&\Longrightarrow (U^TU)^{-1}U^TUSV^TV(V^TV)^{-1} &&= (U^TU)^{-1}U^TXV(V^TV)^{-1}
\end{aligned}$$

$$\therefore \quad S = (U^TU)^{-1}U^TXV(V^TV)^{-1}.$$

$$\begin{aligned}
\nabla_Q\mathcal{F}_{11} &= M - VQ^T &&= 0\\
&\Longrightarrow VQ^T &&= M\\
&\Longrightarrow V^TVQ^T &&= V^TM\\
&\Longrightarrow (V^TV)^{-1}V^TVQ^T &&= (V^TV)^{-1}V^TM\\
&\Longrightarrow ((V^TV)^{-1}V^TVQ^T)^T &&= ((V^TV)^{-1}V^TM)^T
\end{aligned}$$

Figura 9 – Ilustração do problema WC-FNMTF por meio de fatoração de matrizes. As células com a cor azul-escuro caracterizam valores altos na matriz. As células com a cor azul-claro caracterizam valores baixos na matriz. $X \approx USV^*$, tendo $V^* = V^T$, e $M \approx V^*Q^T$, tendo $V^* = V$. As matrizes U e V são binárias, conforme definição do problema.



$$\therefore \quad Q \quad = \quad ((V^TV)^{-1}V^TM)^T.$$

Desta forma, semelhante aos algoritmos que possuem regras iterativas de atualização (*k-means*, FNMTF e BinOvNMTF), é possível deduzir um algoritmo para o problema WC-FNMTF. O algoritmo 9 detalha essas regras. Para apoiar o uso das regras de minimização do problema $\mathcal{F}_{11}$, algumas propriedades de matrizes foram utilizadas e foram listadas no apêndice C.

2.4 Medidas de validação de agrupamento

O processo de agrupamento é uma tarefa de aprendizado não supervisionado e desafiadora no que diz respeito a encontrar um número ótimo de grupos. Mecanismos de validação de um modelo de agrupamento, também conhecido como *Cluster Validity*, oferecem alternativas para ajudar na decisão sobre o número adequado de grupos. As estratégias de validação consideram três abordagens distintas: validação baseada em i) critérios externos, ii) critérios internos e iii) critérios relativos (THEODORIDIS; KOUTROUMBAS, 2008).

Algoritmo 9 Algoritmo WC-FNMTF - Fatoração Tripla Rápida de Matrizes Não Negativas Regularizada com Coocorrência de Palavras

function WC-NMTF(X, k, l, C, itr^{max})
  Initialize: $U^{(0)} \leftarrow 0,1 \mid \sum_{p=1}^{k} u_{ip} = 1$, $V^{(0)} \leftarrow 0,1 \mid \sum_{q=1}^{l} v_{jq} = 1, \forall i, j$, $S^{(0)} \leftarrow \mathcal{U}(0,1)$, $Q^{(0)} \leftarrow \mathcal{U}(0,1)$, $\lambda \leftarrow 1$, $M \leftarrow generateSPPMI(C)$ e $t \leftarrow 0$.
  while $(t \leq t^{max})$ e (não convergiu) **do**

$$S^{(t+1)} \leftarrow (U^{(t)^T} U^{(t)})^{-1} U^{(t)^T} X V^{(t)} (V^{(t)^T} V^{(t)})^{-1}$$

$$Q^{(t+1)} \leftarrow ((V^{(t)^T} V^{(t)})^{-1} V^{(t)^T} M)^T$$

$$\widetilde{V} \leftarrow S^{(t+1)} V^{(t)^T}$$

$$(U^{(t+1)})_{ip} \leftarrow \begin{cases} 1 & p = \arg\min_{p' \in \{1,\ldots,k\}} \|\mathbf{x}_{i\cdot} - \widetilde{\mathbf{v}}_{p'\cdot}\|^2 \\ 0 & \textit{caso contrário} \end{cases} \quad \forall i, p$$

$$\widetilde{U} \leftarrow U^{(t+1)} S^{(t+1)}$$

$$(V^{(t+1)})_{jq} \leftarrow \begin{cases} 1 & q = \arg\min_{q' \in \{1,\ldots,l\}} \|\mathbf{x}_{\cdot j} - \widetilde{\mathbf{u}}_{\cdot q'}\|^2 \\ 0 & \textit{caso contrário} \end{cases} \quad \forall j, q$$

    $t \leftarrow t + 1$
  end while
  return $U^{(t)}, S^{(t)}, V^{(t)}, Q^{(t)}$
end function

Fonte: Waldyr Lourenço de Freitas Junior, 2023

A primeira delas é aquela cujos resultados de agrupamento são comparados a uma estrutura externa preestabelecida. Essa estrutura expressa uma percepção *a priori* sobre a estrutura dos grupos esperada. A segunda é aquela cujos resultados são avaliados em termos quantitativos envolvendo os próprios dados. Por fim, a terceira abordagem objetiva avaliar a estrutura de um agrupamento comparando-a com outros esquemas de agrupamento decorrentes de execuções do mesmo algoritmo, mas com diferentes parâmetros. A compacidade e a separabilidade dos grupos são critérios propostos para validação deles.

Nas seções a seguir serão descritos os dois índices que foram utilizados neste trabalho. O primeiro deles é o Índice de Rand Ajustado, um índice de validação baseada em critérios externos, muito comum na literatura de agrupamento. O outro é o índice *Silhouette*, cuja validação é baseada em critérios internos. Outros índices também encontrados na literatura de agrupamento não serão detalhados nesta seção, mas podem ser encontrados no trabalho de Desgraupes (2013).

2.4.1 Índice de Rand Ajustado

O índice de Rand (RI) (RAND, 1971) é considerado uma medida bem difundida para validação de agrupamento. Pode ser visto como um índice de validação externo, pois considera uma estrutura previamente conhecida para comparar com os grupos descobertos.

Formalmente, seja X uma matriz de dados, $C = \{c_1, \ldots, c_k\}$ a estrutura do agrupamento dos dados com k grupos e $P = \{p_1, \ldots, p_q\}$ uma partição dos dados em X, conhecida antecipadamente, com q partes. Os pares de elementos do conjunto de dados podem ser definidos como (x_p, x_c), sendo $p \neq c$ e $(x_p, x_c) = (x_c, x_p)$. As seguintes afirmações podem ser feitas para cada um dos pares:

- SS: Se ambos os pontos pertencem ao mesmo grupo da estrutura de grupos C e à mesma parte na partição P.
- SD: Se pontos pertencem ao mesmo grupo da estrutura de grupos C e a diferentes partes na partição P.
- DS: Se pontos pertencem a diferentes grupos da estrutura de grupos de C e à mesma parte na partição P.
- DD: Se ambos os pontos pertencem a diferentes grupos da estrutura de grupos de C e a diferentes partes na partição P.

Assume-se que a, b, c e d são as quantidades de pares em SS, SD, DS e DD, respectivamente. A quantidade total dos pares não ordenados em um conjunto de n elementos é dado por $\binom{n}{2}$, que é o mesmo que a soma de a, b, c e d. Com essas definições é possível determinar o índice RI na equação 12.

$$RI = \frac{(a+d)}{(a+b+c+d)} \tag{12}$$

O índice possui uma variação entre 0 e 1, sendo 0 quando os grupos descobertos são totalmente inconsistentes ($a = d = 0$) e 1 quando os grupos descobertos são totalmente consistentes ($b = c = 0$).

O índice de Rand ajustado foi proposto no trabalho de Hubert e Arabie (1985) em que os autores determinaram o valor esperado do índice. O índice de Rand original não é ajustado para corrigir o acaso, ou seja, o valor esperado não é nulo para duas partições completamente aleatórias. Dado um conjunto de n objetos $S = \{O_1, \ldots, O_n\}$, e C e P

definidos anteriormente, temos que $\cup_{i=1}^{q} p_i = S = \cup_{j=1}^{k} c_j$, e $u_i \cap u_{i'} = \oslash = v_j \cap v_{j'}$ para $1 \leq i \neq i^{'} \leq q$ e $1 \leq j \neq j^{'} \leq k$. O número de objetos na partição e nos grupos é fixo. Seja n_{ij} o número de objetos que estão na parte p_i e no grupo c_j. Seja n_{i*} e n_{*j} o número de objetos na parte p_i e no grupo c_j, respectivamente, conforme ilustrado na Tabela 1.

Tabela 1 – Tabela de contingência para comparar duas partições

Partição/Grupo	C_1	C_2	...	C_k	*Soma*
P_1	n_{11}	n_{12}	...	n_{1k}	n_{1*}
P_2	n_{21}	n_{22}	...	n_{2k}	n_{2*}
⋮	⋮	⋮		⋮	⋮
P_q	n_{q1}	n_{q2}	...	n_{qk}	n_{q*}
Soma	n_{*1}	n_{*2}	...	n_{*k}	$n_{**} = n$

Fonte: Adaptado de Yeung e Ruzzo (2001)

A forma geral de um índice com o valor esperado constante é $\frac{\textit{índice - índice esperado}}{\textit{índice máximo - índice esperado}}$, que possui um limite máximo de 1 e recebe o valor 0 quando o índice é igual ao valor esperado. O índice de Rand ajustado (ARI) proposto no trabalho de Hubert e Arabie (1985) tem a forma apresentada na equação 13.

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i*}}{2} \sum_j \binom{n_{*j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i*}}{2} + \sum_j \binom{n_{*j}}{2}\right] - \left[\sum_i \binom{n_{i*}}{2} \sum_j \binom{n_{*j}}{2}\right] / \binom{n}{2}} \tag{13}$$

2.4.2 Índice *Silhouette*

O índice *Silhouette* é comumente utilizado no processo de validação de agrupamentos (ROUSSEEUW, 1987). Este índice pode ser visto como um índice de validação interno, já que os resultados dos algoritmos são avaliados por meio de informações dos próprios dados submetidos ao agrupamento. Duas informações são necessárias para o cálculo do índice: a estrutura de agrupamento obtida por algum algoritmo e a lista de todas as distâncias entre os objetos. A distância utilizada neste trabalho para o cálculo do índice foi a euclidiana.

Seja X uma matriz de dados, i um objeto qualquer nesta matriz e A o grupo no qual o objeto foi alocado após o processo de agrupamento. Quando A contiver outros objetos além do i, o seguinte cálculo é feito (A figura 10 contém a ilustração):

$a(i)$ = distância média do objeto i a todos os demais objetos dentro de A (Na figura 10, é o tamanho médio de todas as linhas dentro de A).

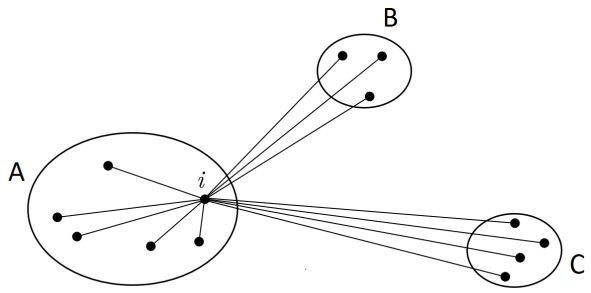
Também, seja C um grupo na estrutura do agrupamento, diferente do grupo A, o seguinte cálculo é feito:

$d(i, C)$ = distância média do objeto i a todos os objetos desse grupo C (Na figura 10, é o tamanho médio de todas as linhas com origem no objeto i até todos os objetos do grupo C.)

Após realizar o cálculo $d(i, C)$ para todos os grupos $C \neq A$, o resultado mínimo é escolhido, denotado por $b(i)$ (assume-se que o número de grupos k é maior que um):

$b(i) = MIN_{C \neq A}\, d(i, C)$.

Figura 10 – Ilustração dos elementos envolvidos no cálculo de $I_s(i)$, em que o objeto i pertence ao grupo A



Fonte: Adaptado de Rousseeuw (1987)

No exemplo ilustrativo da figura 10, o grupo B é chamado de grupo vizinho do objeto i, visto que é o grupo mais próximo dele. Pode-se dizer que o grupo B é o segundo melhor grupo para alocar o objeto i, depois de A. Combinando os cálculos de $a(i)$ e $b(i)$ definidos anteriormente, o índice $I_s(i)$ do objeto i é obtido conforme exposto na equação 14 e uma fórmula pode ser escrita para $I_s(i)$, conforme equação 15.

$$I_s(i) = \begin{cases} \text{I)} & 1 - \frac{a(i)}{b(i)}, \text{ se } a(i) < b(i) \\ \text{II)} & 0, \qquad\quad \text{se } a(i) = b(i) \\ \text{III)} & \frac{b(i)}{a(i)} - 1, \text{ se } a(i) > b(i) \end{cases} \tag{14}$$

$$I_s(i) = \frac{b(i) - a(i)}{MAX\{a(i), b(i)\}} \tag{15}$$

As definições das equações 14 e 15 permitem visualizar que $-1 \leq I_s(i) \leq 1$ e interpretar que:

- Quando $I_s(i)$ atinge o valor máximo de 1 significa que a dissimilaridade no interior do grupo $(a(i))$ é bem menor que a menor dissimilaridade entre grupos $(b(i))$ (I), portanto, o objeto i está bem agrupado ou alocado no grupo correto.

- Quando $I_s(i)$ se aproxima ou é igual a zero (II), significa que $a(i)$ e $b(i)$ são aproximadamente iguais e não está muito evidente a qual grupo o objeto i deveria pertencer, se ao A ou ao B.
- Quando $I_s(i)$ atinge o valor mínimo de -1, significa que a dissimilaridade no interior do grupo ($a(i)$) é bem maior que a menor dissimilaridade entre grupos ($b(i)$) (III), portanto, o objeto i está mal agrupado ou alocado no grupo incorreto.

O $I_s(i)$ é calculado para cada objeto i e o I_s de um grupo é a média dos I_s de todos os objetos (i) de um determinado grupo. O I_s do agrupamento é a média dos I_s de todos os grupos.

3 Estado da arte

Este capítulo objetiva apresentar uma revisão bibliográfica sobre estudos que aplicaram métodos de agrupamento e coagrupamento baseados em fatoração de matrizes em textos. Na seção 3.1 são apresentadas as estratégias para análise qualitativa encontradas na literatura. Na seção 3.2 é apresentado como cada estudo abordou as estratégias de agrupamento e coagrupamento baseadas em fatoração de matrizes, técnica base dos algoritmos estudados neste trabalho de pesquisa. Também é feita uma discussão dos estudos que usaram tais abordagens. A seção 3.3 apresenta os tipos de conjuntos de dados mais comuns utilizados nos estudos. A seção 3.4 apresenta os modelos de representação vetorial encontrados nos estudos. Na seção 3.5 são expostas as estratégias dos estudos para análise quantitativa, ou seja, para validação da capacidade de agrupamentos dos algoritmos. Por fim, a seção 3.6 faz uma rápida discussão dos estudos que compararam os resultados dos métodos propostos com os resultados dos métodos de agrupamento. O quadro 1 apresenta de forma consolidada as discussões das seções apresentadas.

3.1 Estratégias para análise qualitativa

Análise qualitativa realizada sobre textos é um assunto muito subjetivo e tem como finalidade, no contexto deste trabalho, abstrair conhecimento sobre os textos de um corpus, a partir dos resultados dos algoritmos de agrupamento e coagrupamento aplicados sobre esse corpus. Essa análise está ligada à descoberta das relações entre tópicos e palavras subjacentes aos textos sob análise. Essa não é uma tarefa trivial, portanto, alguns estudos sugerem abordagens específicas para produzir algum significado dos resultados dos algoritmos. As estratégias encontradas nos estudos analisados podem ser organizadas da seguinte forma:

- **Análise dos vetores protótipos** (LEE; SEUNG, 1999; CHEN; WANG; DONG, 2009; ALLAB; LABIOD; NADIF, 2016; AILEM; AGHILES; NADIF, 2017; SALAH; AILEM; NADIF, 2018; CASALINO *et al.*, 2018; HASSANI; AMIR; MANSOURI, 2021; FEBRISSY *et al.*, 2022). Após a execução do processo de fatoração de uma matriz de dados, seja uma fatoração dupla ou tripla, é possível encontrar os vetores protótipos que indicam a qual grupo (de linha ou coluna) um dado ou atributo

Quadro 1 – Quadro comparativo de estudos a respeito de agrupamento e coagrupamento de textos baseados em fatoração de matrizes

Estudo	**Conjunto de Dados**	**Medida**	**Abordagem fatoração**	**Repres. Vetorial**	**Análise Qualitativa**	**Agrup. Clássico?**
Febrissy *et al.* (2022)	CSTR, CLASSIC4, RCV1, NG20	ARI, NMI	Dupla	BoW (TF-IDF)	Vet. Protótipo	Sim
Hassani, Amir e Mansouri (2021)	BBC News, BBC Sport, WebACE, NG20	Pureza, ARI, NMI	Dupla	BoW (TF-IDF)	Vet. Protótipo Nuvem de palavras	Sim
Freitas Jr. *et al.* (2020)	NIPS, próprio	ARI	Tripla	BoW (TF, TF-IDF)	Matriz S, Nuvem de palavras	Sim
Abe e Yadohisa (2019)	TREC, Reuters, fbis, hitech, WebACE	Acurácia, ARI	Tripla	BoW (TF-IDF)	Matriz S	Não
Liu, Hua e Chen (2019)	CSTR, CLASSIC4, LA Times, NG20	ARI, NMI	Dupla	*word embedding*	-	Não
Guo *et al.* (2019)	WS-sim, Simlex-999, WS-rel, MEN	Davies Bouldin	Dupla	*word embedding*	Nuvem de palavras	Não
Casalino *et al.* (2018)	Twitter	NMI, *Silhouette*	Dupla	BoW (TF-IDF)	Vet. Protótipo Nuvem de palavras	Sim
Salah, Ailem e Nadif (2018)	CSTR, CLASSIC4, LA Times, NG20	ARI, NMI	Tripla	BoW (TF-IDF)	Vet. Protótipo	Sim
Ailem, Aghiles e Nadif (2017)	CLASSIC4, Reuters, NG20	ARI, NMI	Dupla	*word embedding*	Vet. Protótipo	Sim
Brunialti *et al.* (2017)	NIPS, próprio	ARI	Tripla	BoW (TF, TF-IDF)	-	Sim
Shahid *et al.* (2017)	Twitter	-	Dupla	BoW (TF-IDF)	Nuvem de palavras	Sim
Alzahrani *et al.* (2016)	Próprio	Análise de especialista	Dupla	BoW (TF-IDF)	Análise de especialista	Não
Allab, Labiod e Nadif (2016)	CSTR, CLASSIC4, NG20, WebACE	Acurácia, ARI, NMI	Tripla	BoW (TF-IDF)	Vet. Protótipo	Sim
Yoo e Choi (2010)	CSTR, WebACE, Reuters	Acurácia, NMI	Tripla	BoW (TF-IDF)	Matriz S	Não
Chen, Wang e Dong (2009)	Reuters, WebACE, NG20	Acurácia, MI	Tripla	BoW (TF)	Vet. Protótipo	Não
Lee e Seung (1999)	Enciclopédia Grolier	-	Dupla	BoW (TF)	Vet. Protótipo	Não

Fonte: Waldyr Lourenço de Freitas Junior, 2023

pertence. Baseados nos fatores (valores numéricos) mais representativos de cada vetor protótipo, os estudos propõem discussões sobre os documentos e as palavras que são representados por tais vetores.

- **Análise da matriz** S (YOO; CHOI, 2010; ABE; YADOHISA, 2019; FREITAS JR. *et al.*, 2020). Após a execução do processo de fatoração em um modelo de fatoração tripla, uma das matrizes geradas é a matriz S ($X \approx USV^T$). Essa matriz expressa uma relação entre os grupos de linhas e os grupos de colunas indicados pelas matrizes U e V, respectivamente. Os estudos que propõe explorar essa matriz usam os valores de suas células como informação que relaciona documentos e palavras e indicam polissemia, polarização e tópicos.
- **Visualização de representações textuais dos grupos e cogrupos** (SHAHID *et al.*, 2017; CASALINO *et al.*, 2018; GUO *et al.*, 2019; FREITAS JR. *et al.*, 2020; HASSANI; AMIR; MANSOURI, 2021). Após a execução dos algoritmos, as palavras

associadas aos maiores fatores da matriz V (para palavras que representam os grupos de colunas) e da matriz SV^T (para palavras que mais bem representam os grupos das linhas) são plotadas, por exemplo na forma de nuvens de palavras, para análise de um especialista. Com base nos estudos avaliados, normalmente o próprio autor da pesquisa faz a análise e discussão dessas representações.

Foi observado neste levantamento bibliográfico que existe uma sinergia entre as estratégias acima. Alguns autores utilizam mais de uma delas para realizar análise qualitativa. Observou-se também que os estudos mais antigos utilizaram apenas uma estratégia para análise dos resultados. Com exceção do estudo de Casalino *et al.* (2018), que mesclou a análise dos vetores protótipos com a avaliação de nuvens de palavras, os estudos que realizaram a análise dos vetores protótipos (LEE; SEUNG, 1999; CHEN; WANG; DONG, 2009; ALLAB; LABIOD; NADIF, 2016; AILEM; AGHILES; NADIF, 2017; SALAH; AILEM; NADIF, 2018; HASSANI; AMIR; MANSOURI, 2021; FEBRISSY *et al.*, 2022) não utilizaram mais nenhuma das estratégias supracitadas. Uma das possíveis razões é que a análise de vetores protótipos também pode ser utilizada para validação da qualidade do agrupamento e o principal objetivo dos estudos não era prover uma análise qualitativa dos resultados e sim fazer uma avaliação de desempenho do algoritmo em relação à qualidade dos grupos mediante medidas quantitativas. Em geral, a análise qualitativa é apenas um detalhe a mais na avaliação provida pelos autores.

Uma vantagem da fatoração tripla de matrizes é a geração da matriz fatorada S. A estrutura dessa matriz permite uma análise simultânea da relação entre os grupos de linhas e os grupos de colunas. Essa estratégia mostra-se adequada para contextos nos quais os grupos de dados e grupos de atributos possuem alguma associação, que é o caso dos dados diádicos. Dos estudos analisados, o precursor no uso dessa estratégia foi o de Yoo e Choi (2010). Esse estudo apresentou um novo método de atualização multiplicativa para NMF e NMTF, e discutiu como utilizar a estrutura da matriz S para análise qualitativa dos textos. Os experimentos dessa estratégia demonstraram ser possível detectar polissemia e constatar que palavras mais frequentes nos grupos de palavras identificam tópicos nos documentos.

Outros estudos mais recentes também usaram a matriz S (ABE; YADOHISA, 2019; FREITAS JR. *et al.*, 2020) como estratégia para análise qualitativa. O estudo de Abe e Yadohisa (2019) fez uma discussão detalhada da matriz S sob o conjunto WebACE,

comparando o seu método de análise com aquele proposto no trabalho de Yoo e Choi (2010). O estudo de Freitas Jr. *et al.* (2020) também fez uma discussão detalhada da matriz S aplicada sob um contexto de notícias. A particularidade deste estudo é que o método proposto possui uma matriz V específica de grupos de colunas para cada grupo de linhas. Nessa pesquisa, é defendido que essa estratégia permite maior flexibilidade na geração dos grupos e descrições mais assertivas para cada grupo de documentos.

Os autores que usaram a estratégia de visualização de nuvens de palavras (SHAHID *et al.*, 2017; CASALINO *et al.*, 2018; GUO *et al.*, 2019; FREITAS JR. *et al.*, 2020), fizeram sua própria interpretação dos resultados. O trabalho de Hassani, Amir e Mansouri (2021) fez uso dessa estratégia com o objetivo de demonstrar a capacidade de NMF em separar as palavras em grupos, mas não fez uma discussão sobre o resultado dessa visualização.

De todos os estudos avaliados, apenas o estudo de Alzahrani *et al.* (2016) realizou a análise qualitativa dos grupos com apoio de um especialista no domínio do corpus (textos extremistas islâmicos). Eles geraram seis grupos com a técnica proposta e outros seis grupos com técnicas tradicionais de coagrupamento. Todos os 12 grupos foram apresentados ao especialista para avaliação. Entretanto, este estudo não aplicou uma medida de validação da análise do especialista e nenhuma das estratégias discutidas nesta seção.

3.2 Abordagens de agrupamento e coagrupamento baseadas em fatoração de matrizes

Este trabalho tem interesse nas estratégias de agrupamento e coagrupamento baseados em fatoração de matrizes. Especificamente sobre fatoração de matrizes, na literatura observaram-se abordagens baseadas em fatoração dupla e fatoração tripla, com definições gerais a seguir:

- As **abordagens baseadas em fatoração dupla** são aquelas que utilizaram o NMF tradicional (cf. seção 2.1), voltadas ao agrupamento unilateral.
- As **abordagem baseadas em fatoração tripla** são aquelas que utilizaram o BVD e suas variações (cf. seção 2.3), voltadas ao agrupamento bilateral.

Nas seções 3.2.1 e 3.2.2 são explanados os estudos que utilizam cada uma dessas abordagens.

3.2.1 Abordagens baseadas em fatoração dupla

NMF foi proposta inicialmente como alternativa a métodos tradicionais, como Análise de Componentes Principais e Quantização Vetorial, sobretudo pela sua distinção de restrições de não negatividade (LEE; SEUNG, 1999; LEE; SEUNG, 2000). O trabalho de Lee e Seung (1999) propôs NMF como um método de extração de recursos de imagens e de análise de características semânticas em textos, considerando para esta última uma lista de corpus de artigos de enciclopédia. Entretanto, NMF é uma abordagem que usa a fatoração dupla de matrizes e não gera grupos de colunas durante o processo de fatoração do algoritmo. Segundo o estudo de Yoo e Choi (2010), etapas de pós processamento são necessárias para realizar uma correspondência entre documentos e palavras, e necessárias para determinar polissemia em documentos que pertençam a diferentes contextos. De uma forma geral, os trabalhos que se basearam na abordagem de fatoração dupla, utilizaram-na para atingir objetivos relacionados à avaliação quantitativa.

O estudo de Ailem, Aghiles e Nadif (2017) propôs um novo algoritmo de NMF com *word embedding*, pressupondo preservar a relação semântica entre as palavras.

O estudo de Alzahrani *et al.* (2016) utilizou uma versão otimizada de NMF, proposto no trabalho de Li e Ngom (2013), para processar relações sujeito-verbo-objeto. O objetivo foi detectar tais relações baseadas em unigramas e bigramas, e compará-las quando baseadas em conceitos. O estudo de Shahid *et al.* (2017) utilizou uma versão de NMF da biblioteca *scikit-learn* do Python. O estudo propôs analisar tendências extremistas do Twitter[1] comparando com outros métodos.

O estudo de Casalino *et al.* (2018) utilizou quatro variações de NMF e explorou a forma de inicialização de cada uma delas. O objetivo foi melhorar a qualidade do resultado do agrupamento, já que NMF é sensível à inicialização. As quatro variações foram aplicadas sobre dados do Twitter em busca de descoberta de tópicos. O estudo de Guo *et al.* (2019) usou o NMF tradicional para decompor uma matriz de coocorrência de palavras para duas novas matrizes. A matriz que representa os vetores protótipos serviu de entrada para um novo modelo baseado em *fuzzy k-means*, cujo propósito era detecção de similaridade e relações entre palavras.

O estudo de Liu, Hua e Chen (2019) tratou a questão de preservação semântica por meio da coocorrência de palavras sob um modelo de fatoração dupla pré-existente, o

[1] Twitter é uma rede social e um serviço de microblog. https://twitter.com/

DNMF (*Graph Dual Regularization Non-negative Matrix Factorization*). Ele apresenta uma abordagem semelhante ao estudo de Salah, Ailem e Nadif (2018), mas a principal diferença é que este último usou fatoração tripla. A extensão do método consistiu em incorporar dois termos de regularização à DNMF, um para regularização dos documentos (dados) e outro para regularização das palavras (atributos).

O estudo de Hassani, Amir e Mansouri (2021) utilizou NMF como um passo intermediário para melhorar o agrupamento. A proposta foi combinar características durante o processo de fatoração das matrizes a fim de criar um espaço de características menor. Em seguida, os novos vetores são submetidos ao processo de agrupamento pelo *k-means*. Eles também propuseram utilizar Decomposição em Valores Singulares (SVD) como mecanismo de inicialização de NMF. Segundo os autores, este método gera agrupamentos mais adequados.

Um estudo mais recente proposto por Febrissy *et al.* (2022) objetiva melhorar o agrupamento de NMF por meio da utilização do relacionamento das palavras nos corpora. O método assume que as palavras que aparecem no mesmo contexto, por exemplo, na mesma frase ou no mesmo documento, possuem alguma relação semântica entre si. O estudo demonstra por meio de experimentos que essa relação semântica aumenta fortemente o desempenho do agrupamento de NMF.

3.2.2 Abordagens baseadas em fatoração tripla

O trabalho de Long, Zhang e Yu (2005) propôs uma nova estrutura de agrupamento e coagrupamento de dados, intitulada Decomposição de Valores em Blocos Não Negativos (NBVD), também conhecida como NMTF. Essa estrutura pode ser vista como uma extensão de NMF, contudo, em vez de decompor a matriz original em duas novas matrizes, NMTF decompõe a matriz original em três novas matrizes.

O estudo de Chen, Wang e Dong (2009) apresentou um método que adapta NMF e o chama de semi-supervisionado. Tal método estima novas matrizes *palavra-documento* e *documento-categoria*, incorporando restrições cedidas pelos usuários. Esse conhecimento prévio fornecido pelos usuários especifica se um documento deve (*must-link*) ou não deve (*cannot-link*) ser agrupado, melhorando assim a qualidade do coagrupamento.

O método proposto no estudo de Yoo e Choi (2010) propõe preservar as restrições de ortogonalidade das matrizes durante o processo de atualização por meio da *Variedade de Stiefel*. O trabalho propôs essa abordagem em contraste à abordagem proposta no trabalho de Ding *et al.* (2006), que se baseou nos multiplicadores de Lagrange para atualização de NMF e NMTF. A justificativa é que esse método preserva mais eficientemente a ortogonalidade durante o processo de fatoração, o que melhora o desempenho do agrupamento de documentos.

O estudo de Allab, Labiod e Nadif (2016) propôs um modelo híbrido com SemiNMF e PCA. SemiNMF relaxa as restrições de não negatividade e permite que a matriz X e a matriz indicadora de grupos tenham sinais mistos. PCA permite encontrar o subespaço ótimo de baixa dimensionalidade. A proposta dos autores foi melhorar a precisão do agrupamento dos dados. O estudo de Salah, Ailem e Nadif (2018) propôs um novo método que incorpora um termo de regularização sobre NMTF. O novo método, intitulado WC-NMTF, é fundamentado em uma matriz de coocorrência de palavras que se baseia na informação mútua pontual (*Point-wise Mutual Information* - PMI). Dessa forma, assumindo que PMI é altamente correlacionada com os julgamentos humanos para relações entre palavras, o novo método propôs melhorar o desempenho do NMTF clássico, tanto em termos de agrupamento das palavras quanto de agrupamento dos documentos.

Os métodos propostos no trabalho de Abe e Yadohisa (2019) são uma extensão dos métodos ONMTF propostos nos trabalhos de Yoo e Choi (2010) e Ding *et al.* (2006). Os novos métodos são baseados em diferentes distribuições. Os métodos anteriores assumiam que a distribuição do erro seguia uma distribuição Normal, entretanto, os autores contrapõem essa premissa argumentando que isso nem sempre acontece para dados não negativos. Segundo os autores, os métodos propostos facilitam a interpretação dos grupos de linhas e colunas, e são mais robustos com valores extremamente grandes.

Por fim, os estudos de Brunialti *et al.* (2017) e Freitas Jr. *et al.* (2020) apresentaram novos métodos baseados em NMTF para tratar o problema de sobreposição de colunas, os algoritmos intitulados BinOvNMTF e OvNMTF, respectivamente. Tais algoritmos geram k diferentes matrizes indicadoras dos grupos de colunas, em que k é o número de grupos de linhas desejado. Ainda assim, a essência do NMTF é mantida, pois, são geradas três tipos de matrizes distintas (U, S e V). As principais diferenças dos algoritmos podem ser analisadas nas seções 2.3.3 e 2.3.5.

3.3 Dados textuais

Os experimentos realizados para avaliar a capacidade de agrupamento de um algoritmo podem ser executados sobre conjuntos de dados sintéticos ou sobre conjuntos de dados reais. Os experimentos para extrair conhecimento para uma análise qualitativa dos resultados dos algoritmos têm sido realizados sob conjunto de dados reais. Não foi encontrado nenhum estudo que criasse um ambiente controlado, a partir de um conjunto de dados real ou de um conjunto de dados sintéticos, sobre o qual fosse possível realizar uma análise qualitativa de extração de conhecimento.

A maior parte dos estudos analisados utilizaram conjuntos de dados textuais clássicos (CHEN; WANG; DONG, 2009; YOO; CHOI, 2010; ALLAB; LABIOD; NADIF, 2016; AILEM; AGHILES; NADIF, 2017; SALAH; AILEM; NADIF, 2018; ABE; YADOHISA, 2019; HASSANI; AMIR; MANSOURI, 2021; FEBRISSY *et al.*, 2022). Esses conjuntos de dados possuem características peculiares para validação de agrupamento, como diferentes números de grupos, diversos domínios, quantidade variada de dados, diferentes graus de sobreposição de grupos e, na maior parte, são rotulados. Não foram encontrados conjuntos de dados para avaliar a qualidade dos grupos de palavras. Essa tarefa permanece sendo um desafio, visto que os conjuntos de referência não viabilizam os rótulos para as palavras (SALAH; AILEM; NADIF, 2018).

Os conjuntos de dados reais utilizados nos estudos discutidos nesta seção estão apresentados no quadro 1. Os conjuntos mais utilizados estão detalhados a seguir:

- **CLASSIC4**[2]: Coleção composta por quatro conjuntos distintos: CACM (resumos de artigos da ACM), CISI (artigos de recuperação de informações), CRANFIELD (artigos de sistema aeronáutico) e MEDLINE (periódicos da área médica).
- **CSTR**[3]: Coleção de relatórios técnicos de ciências da computação da Universidade de Rochester, publicados entre 1991 e 2007. As principais categorias do conjunto são: Processamento de Linguagem Natural (NLP), robótica e visão computacional.
- **LA Times**: Coleção formada por artigos de notícias do *Los Angeles Times* extraídos do TREC[4].

[2] http://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets/
[3] https://www.cs.rochester.edu/research/technical_reports.html
[4] http://trec.nist.gov/data.html

- **NG20**[5] (20 *Newsgroups*): Coleção composta por 20.000 mensagens extraídas do grupo Usenet.
- **Reuters**: Coleção de notícias da agência britânica de notícias Reuters[6].
- **WebACE**: Coleção formada pelas páginas *web* do Yahoo![7].

Outros conjuntos de dados também foram utilizados. Alguns autores utilizaram conjuntos próprios (ALZAHRANI *et al.*, 2016; BRUNIALTI *et al.*, 2017; FREITAS JR. *et al.*, 2020) e outros utilizaram dados do Twitter (CASALINO *et al.*, 2018; SHAHID *et al.*, 2017). Os estudos que utilizaram dados do Twitter objetivaram a análise de discurso de ódio e análise de sentimentos sob determinados tópicos. O estudo de Hassani, Amir e Mansouri (2021) utilizou dois conjuntos de notícias da BBC[8] (BBC News e BBC Sports).

De todos os estudos avaliados, os que realizaram algum experimento com dados sintéticos, o fizeram com conjuntos não textuais criados pelos próprios autores.

3.4 Representação vetorial

Além das etapas tradicionais de pré-processamento de textos, como "tokenização", remoção de "stopwords", aplicação de "stemming", aplicação de "case-folding", etc., para a etapa de processamento, os conjuntos de dados precisam ser organizados em um modelo de espaço vetorial. Nos estudos avaliados foram observadas duas representações vetoriais: *bag-of-words* (BoW) e *word embedding*.

Os métodos tradicionais da abordagem NMF/NMTF utilizam a representação BoW, entretanto, segundo o estudo de Ailem, Aghiles e Nadif (2017), existe uma deficiência nesse tipo de abordagem quando o objetivo do estudo está relacionado ao significado das palavras. Segundo o estudo, as representações BoW não levam em conta a sequência na qual as palavras aparecem em um documento. Essa é uma questão importante, pois pode resultar em perda de informações, especialmente as relações semânticas entre palavras.

Dentre os estudos avaliados, o estudo de Ailem, Aghiles e Nadif (2017) foi o primeiro a contrapor a questão da representação dos textos. Nesse estudo, é proposto um novo método para lidar com o problema de preservação da relação semântica entre palavras, assumindo que as palavras que coocorrem na mesma direção do espaço, ou seja, no mesmo

[5] http://qwone.com/~jason/20Newsgroups/
[6] https://www.reuters.com
[7] http://www.yahoo.com/
[8] https://www.bbc.com/

contexto, possuem significados semelhantes. Essa questão foi abordada integrando o modelo "word2vec" em uma estrutura de NMF e uma matriz de coocorrência das palavras, o que demonstrou ser eficiente em promover as relações semânticas entre as palavras.

Os trabalhos anteriores a esse utilizaram a representação tradicional (BoW), baseados na frequência das palavras (*Term Frequency* - TF[9] e *Term Frequency Inverse Document Frequency* - TF-IDF[10]). Observou-se também que a partir do ano de 2017 outros estudos também passaram a usar *word embedding*[11] como um modelo de representação de palavras.

3.5 Validação do agrupamento

Dentre os estudos analisados, foi possível identificar pelos menos duas abordagens de validação do agrupamento dos dados. A abordagem menos comum nos estudos aborda o problema de agrupamento sem comparar o resultado com uma estrutura externa predefinida. As validações deste tipo de abordagem são feitas por meio de medidas internas, ou seja, que avaliam os resultados em termos quantitativos envolvendo os próprios dados. Entretanto, a abordagem mais comum aborda o problema de agrupamento dos dados de forma semelhante a um problema de classificação, confrontando o resultado dos algoritmos com rótulos reais. As validações deste tipo de abordagem são feitas por meio de medidas consideradas externas, ou seja, que comparam os resultados com uma estrutura externa preestabelecida.

Apenas dois estudos, entre aqueles analisados neste trabalho de pesquisa, abordaram a validação do agrupamento com base em medidas internas. O estudo de Guo *et al.* (2019) usou o índice Davies-Bouldin e o estudo de Casalino *et al.* (2018) utilizou o índice *Silhouette*. Este segundo estudo, dentre todos os analisados, foi o único que utilizou ambas as abordagens acima para avaliar o agrupamento dos dados, além de considerar também o erro de reconstrução das matrizes e o tempo de execução. A estratégia de usar ambas as abordagens pode ser interessante, uma vez que a tarefa de agrupamento é naturalmente não

[9] Modelo de representação de documentos que é definido pelo número de vezes que uma palavra aparece em um documento.

[10] Modelo de representação de documentos que é definido pela multiplicação de duas medidas. A primeira é a TF, já definida, e a segunda é a IDF, que é a fração inversa dos documentos que contêm determinada palavra. Esse modelo mede quão importante uma palavra é em um documento e o quanto de informação ela providencia.

[11] Modelo de representação de palavras, tipicamente por meio de vetores de números reais, de maneira que palavras com significados semelhantes possuem representações similares.

supervisionada e as medidas externas pressupõem que o número de grupos seja conhecido previamente, o que nem sempre é possível.

A maior parte dos estudos utilizou a abordagem baseada em medidas externas. Os índices mais utilizados foram a Informação Mútua Normalizada (NMI) e o Índice de Rand Ajustado (ARI). Uma particularidade identificada nos estudos foi que quase todos os que utilizaram o ARI, o fizeram para validação do agrupamento das linhas. A exceção foi o estudo de Abe e Yadohisa (2019), que também o fez para validação dos grupos de colunas, utilizando conjuntos com dados sintéticos. O único estudo que utilizou a medida Pureza, dentre os aqui apresentados, foi o de Hassani, Amir e Mansouri (2021).

3.6 Comparação com agrupamento clássico

Uma das questões observadas nos estudos analisados foi que alguns deles compararam os métodos propostos ou utilizados com o algoritmo *k-means* e suas variações.

Os estudos que realizaram tal comparação (ALLAB; LABIOD; NADIF, 2016; SHAHID *et al.*, 2017; BRUNIALTI *et al.*, 2017; AILEM; AGHILES; NADIF, 2017; SALAH; AILEM; NADIF, 2018; CASALINO *et al.*, 2018; FREITAS JR. *et al.*, 2020; FEBRISSY *et al.*, 2022), no geral justificaram que existe uma correspondência entre *k-means* e NMF. O problema de agrupamento *k-means* pode ser definido usando fatoração de matrizes (DING; HE, 2005), o que o torna um candidato para comparação com outros métodos baseados em fatoração de matrizes. Além disso, segundo o estudo de Wang *et al.* (2011), ao incluir restrições de ortogonalidade no problema de NMF, os objetivos de NMF se aproximam dos resultados de *k-means*, tornando-os equivalentes. Essas são algumas das razões pelas quais ele tem sido frequentemente usado como referência para experimentos realizados com os métodos baseados em NMF (BRUNIALTI *et al.*, 2017).

Em contrapartida, metade dos estudos analisados não compararam os resultados dos métodos utilizados com aqueles gerados por *k-means* (LEE; SEUNG, 1999; CHEN; WANG; DONG, 2009; YOO; CHOI, 2010; ALZAHRANI *et al.*, 2016; GUO *et al.*, 2019; LIU; HUA; CHEN, 2019; ABE; YADOHISA, 2019). Nenhum desses autores justificou essa tomada de decisão em seus estudos.

4 Experimentos e resultados

Este capítulo apresenta os experimentos realizados e os resultados alcançados com os algoritmos de agrupamento e coagrupamento baseados em fatoração de matrizes. A seção 4.1 apresenta os experimentos realizados com dados sintéticos e a seção 4.2 apresenta os experimentos com dados do mundo real, para avaliação quantitativa. A seção 4.3 apresenta os procedimentos de avaliação qualitativa executados por meio de interação com alunos de graduação em Sistemas de Informação da Universidade de São Paulo. Cada uma dessas seções apresenta uma discussão dos conjuntos de dados, da configuração dos experimentos e dos resultados.

4.1 Experimentos com dados sintéticos

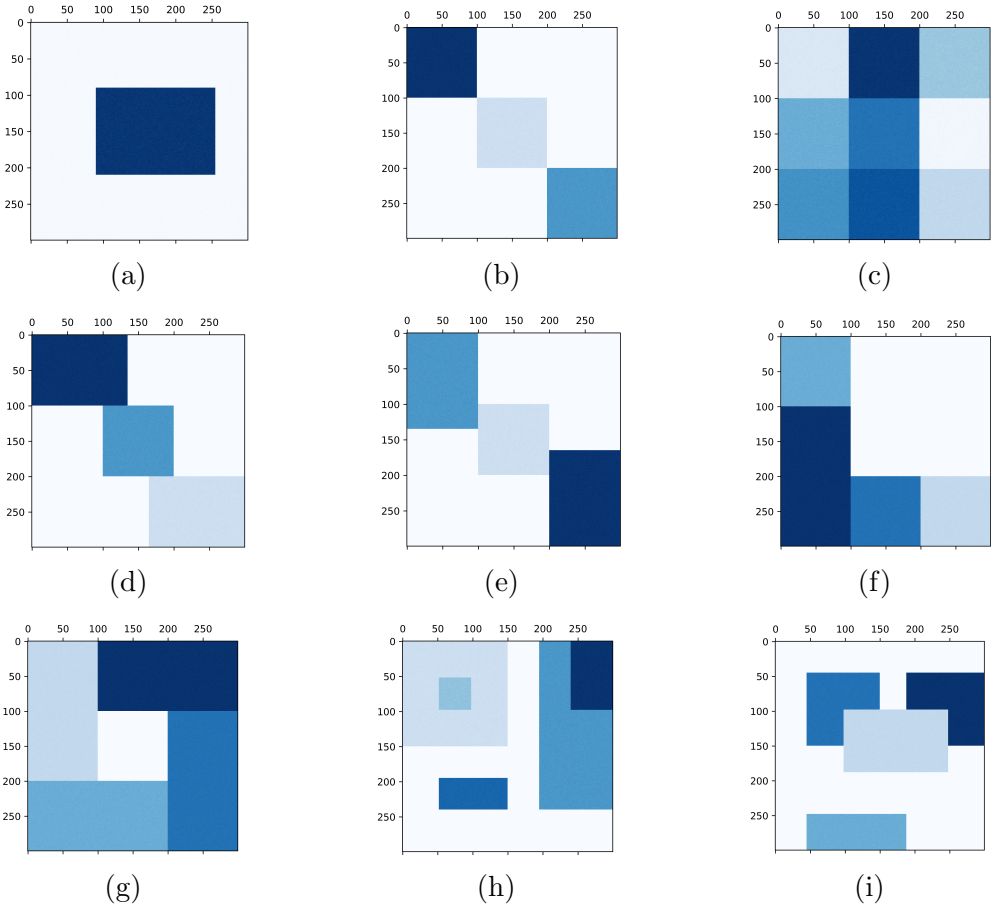
Esta seção apresenta os experimentos realizados com dados sintéticos. Na seção 4.1.1 é explicado como os conjuntos de dados e seus respectivos rótulos foram gerados. A seção 4.1.2 apresenta a configuração dos experimentos. A seção 4.1.3 discute a estratégia utilizada para determinar a pertinência a grupos de linhas e colunas. A seção 4.1.4 apresenta uma discussão dos resultados decorrentes dos experimentos. Por fim, a seção 4.1.5 faz as considerações finais.

4.1.1 Conjuntos de dados

Geração dos dados

A fim de realizar experimentos controlados com relação ao coagrupamento dos dados, foram gerados conjuntos de dados sintéticos com diferentes estruturas de cogrupos (MADEIRA; OLIVEIRA, 2004). A figura 11 apresenta essas estruturas, a saber: 11a único cogrupo, 11b cogrupos com linhas e colunas exclusivas, 11c cogrupos com estrutura de tabuleiro de xadrez, 11d cogrupos com linhas exclusivas, 11e cogrupos com colunas exclusivas, 11f cogrupos sem sobreposição com estrutura em árvore, 11g cogrupos não exclusivos e sem sobreposição, 11h cogrupos com sobreposição e com estrutura hierárquica e 11i cogrupos com sobreposição e arbitrariamente posicionados.

Figura 11 – Estruturas de cogrupos que foram utilizadas para os experimentos



Fonte: Waldyr Lourenço de Freitas Junior, 2023

Para cada estrutura de cogrupos, foi gerada uma matriz X de tamanho $n \times m$. Cada elemento x_{ij} dessa matriz foi gerado aleatoriamente seguindo uma função que gera uma distribuição uniforme $U(0,1) \in]0,1]$. Para gerar os cogrupos dentro da matriz X, foram geradas submatrizes cujas linhas se iniciam em um determinado elemento x_{ij_1} e finalizam em um elemento x_{ij_2} e cujas colunas se iniciam em um determinado elemento x_{i_1j} e finalizam em um elemento x_{i_2j}.

Todos os elementos de uma dessas submatrizes foram regerados aleatoriamente seguindo uma função que gera uma distribuição uniforme $U(0,10) \in]0,10]$. Em seguida, ainda para todos os elementos dessa mesma submatriz, foi somado um valor inteiro, aleatoriamente escolhido dentro de um conjunto C (para cada estrutura foi utilizado um conjunto C distinto, conforme tabela 2). Esse valor foi somado, pois proporciona a diferenciação dos cogrupos, portanto, toda vez que um elemento de C foi escolhido, ele foi em seguida excluído do conjunto. Por essa razão, o número de elementos do conjunto C é igual ao número de cogrupos, para uma estrutura específica.

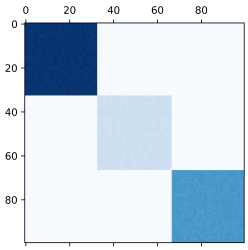
Tabela 2 – Valores definidos para o conjunto C para cada uma das estruturas de cogrupos

Estrutura de Cogrupo	Conjunto C
11a	{150}
11b	{50, 150, 250}
11c	{75, 100, 125, 150, 175, 200, 225, 250, 275}
11d	{50, 150, 250}
11e	{50, 150, 250}
11f	{75, 150, 225, 300}
11g	{75, 150, 225, 300}
11h	{50, 100, 150, 200, 250}
11i	{75, 150, 225, 300}

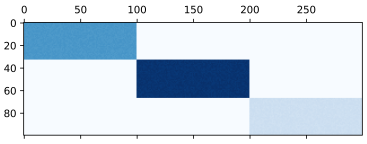
Fonte: Waldyr Lourenço de Freitas Junior, 2023

Especificamente para os experimentos deste trabalho, a fim de trazer mais variedade aos experimentos, os conjuntos gerados foram diversificados de duas maneiras. A primeira delas, por número de linhas (dados) maior, menor ou igual ao número de colunas (atributos). A outra, por um fator de esparsidade que inclui zeros nos elementos da matriz, dado um percentual. Para cada estrutura de cogrupos, foram gerados quatro conjuntos distintos com o tamanho $n \times m$, em que n é 100 ou 300 e m é 100 ou 300. A figura 12 exemplifica a estrutura da figura 11b, com três cogrupos de interesse.

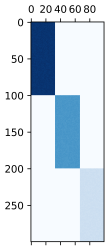
Figura 12 – Ilustração dos diferentes conjuntos gerados para a estrutura da figura 11b (Sem escala). (a) conjunto com tamanho 100 × 100, (b) conjunto com tamanho 100 × 300, (c) conjunto com tamanho 300 × 100 e (d) conjunto com tamanho 300 × 300



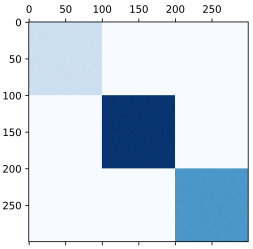
(a) 100 × 100



(b) 100 × 300



(c) 300 × 100

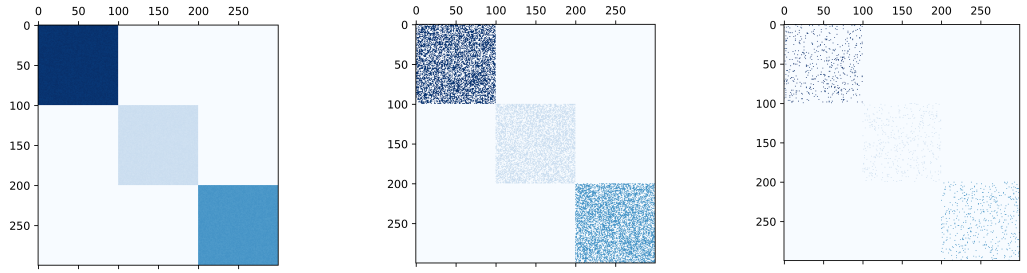


(d) 300 × 300

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Para cada um destes novos conjuntos de dados, foram geradas três instâncias de conjunto: a original, com 0% de esparsidade (dados densos), uma com 50% de esparsidade (fator médio) e uma com 95% de esparsidade (fator altor). Para ilustrar, a figura 13 apresenta a estrutura da figura 11b, 300×300, com fatores de esparsidade 0%, 50% e 95%.

Figura 13 – Ilustração de conjuntos de dados gerados com fatores de esparsidade distintos, baseados na estrutura da figura 11b



(a) Esparsidade 0% (b) Esparsidade 50% (c) Esparsidade 95%

Fonte: Waldyr Lourenço de Freitas Junior, 2023

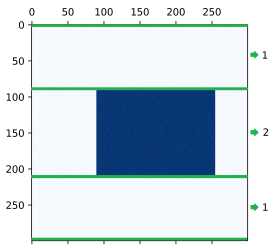
Para cada estrutura, foram gerados doze conjuntos de dados diferentes, seguindo as características explicadas anteriormente. Considerando as nove estruturas, para o experimento como um todo, foram utilizados 108 conjuntos de dados distintos.

Geração dos rótulos de linhas e colunas

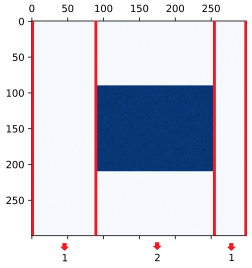
Por tratar-se de um experimento controlado, cada um dos conjuntos de dados sintéticos foi rotulado em relação aos seus grupos de linhas e aos seus grupos de colunas. Foi explanado anteriormente que os cogrupos foram gerados com suas linhas iniciando em um determinado elemento x_{ij_1} e finalizando em um elemento x_{ij_2} e suas colunas se iniciando em um determinado elemento x_{i_1j} e finalizando em um elemento x_{i_2j}. Baseado nessas definições, uma estrutura de dados do tipo vetor de tamanho n com os rótulos de linhas e uma outra estrutura do mesmo tipo e de tamanho m com os rótulos de colunas foram geradas. O rótulo de cada grupo distinto, tanto de linha quanto de coluna, foi enumerado com valores inteiros. Para ilustrar a geração desses rótulos, a figura 14 apresenta as estruturas das figuras 11a e 11e como exemplos.

Na figura 14a observa-se que há dois grupos distintos nas linhas. Uma estrutura de dados do tipo vetor de tamanho 300 foi gerado e cada um dos elementos dessa estrutura foi rotulado como pertencente ao grupo 1 ou ao grupo 2. Nota-se que as primeiras e as últimas linhas do conjunto pertencem ao grupo 1, pois tais linhas foram geradas pela

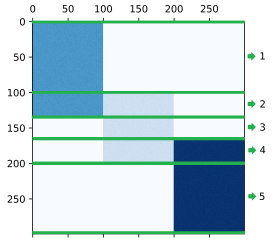
Figura 14 – Ilustração de como foram rotulados os conjuntos de dados sob a ótica dos grupos de linhas e colunas



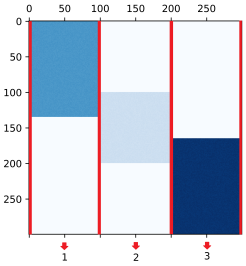
(a) Estrutura da figura 11a rotulada para dois grupos de linhas



(b) Estrutura da figura 11a rotulada para dois grupos de colunas



(c) Estrutura da figura 11e rotulada para cinco grupos de linhas



(d) Estrutura da figura 11e rotulada para três grupos de colunas

Fonte: Waldyr Lourenço de Freitas Junior, 2023

mesma distribuição de dados. A mesma explicação pode ser derivada para a figura 14b. Na figura 14c observa-se que há cinco grupos distintos nas linhas, enquanto na figura 14d, três grupos de colunas são observados. Todos os conjuntos de dados foram rotulados seguindo essa estratégia.

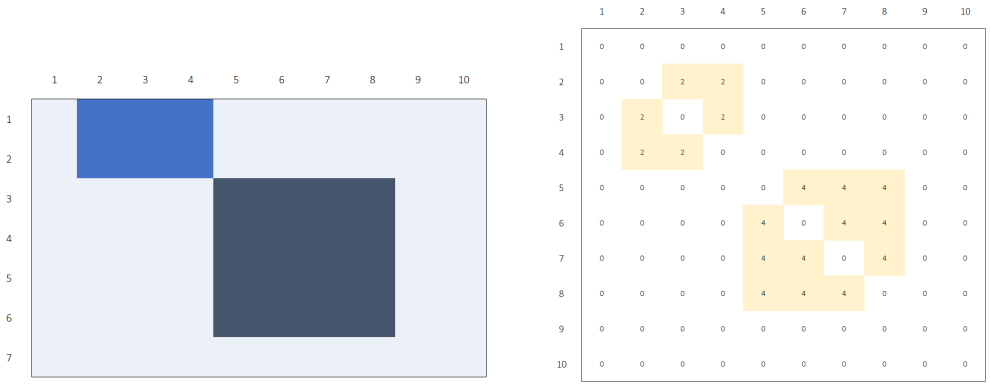
Geração das matrizes de coocorrência

Para a realização dos experimentos com dados sintéticos sobre os algoritmos WC-NMTF e WC-FNMTF, detalhados nas seções 2.3.6 e 2.3.7, respectivamente, foi necessária a construção de uma matriz de coocorrência sintética para cada um dos conjuntos. Essa matriz de coocorrência foi inspirada na ideia de que a coocorrência entre duas palavras está representada nos cogrupos gerados pelos algoritmos; assume-se que esse seja o efeito da coocorrência de palavras em um texto (palavras similares e/ou relacionadas fazem parte da formação de determinado cogrupo). Por exemplo, em uma notícia sobre educação, em que as palavras “ensino” e “qualidade” sempre aparecem no mesmo contexto, elas farão parte de um cogrupo sobre ensino.

Esta seção detalha como foi realizada a construção dessas matrizes.

A figura 15a exemplifica um conjunto de dados sintéticos contendo dois cogrupos, em que o número de linhas é sete e o número de colunas é dez, numeradas na figura. O primeiro cogrupo (c1) compreende as linhas 1 a 2, e as colunas 2 a 4. O segundo cogrupo (c2) compreende as linhas 3 a 6, e as colunas 5 a 8.

Figura 15 – Exemplo de (a) um conjunto de dados com dois cogrupos usado para construir a matriz de coocorrência sintética e (b) da matriz de coocorrência sintética gerada para esse conjunto



(a) (b)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

A matriz de coocorrência será uma matriz $m \times m$, em que m é o número de colunas do conjunto. Para o exemplo da figura 15a, uma matriz de tamanho 10 por 10 será construída. A coocorrência entre duas colunas é estabelecida quando elas estão no mesmo cogrupo. Usando como exemplo a mesma figura 15a, as colunas que coocorrem para a linha 1 do cogrupo c1 são: 2 e 3, 2 e 4, e 3 e 4. O mesmo acontece para a linha 2 do mesmo cogrupo.

A matriz de coocorrência contabilizará uma coocorrência para cada par de colunas e para cada linha. A figura 15b exemplifica a matriz de coocorrência sintética para o conjunto exemplo da figura 15a.

A célula da linha 2 e coluna 3, da matriz da figura 15b, está contabilizando duas coocorrências do conjunto c1, uma para as colunas 2 e 3 da linha 1 e outra para as mesmas colunas para a linha 2. O mesmo é feito para as demais coocorrências.

Para execução dos experimentos, para cada um dos conjuntos gerados, foi gerada uma matriz de coocorrência sintética no padrão aqui detalhado.

4.1.2 Configuração dos experimentos

Os algoritmos utilizados nos experimentos foram *k-means*, NBVD, ONM3F, ONMTF, OvNMTF, FNMTF, BinOvNMTF, WC-NMTF e WC-FNMTF. Para cada um dos conjuntos de dados, supõe-se que o número de grupos de linhas (k) e o número de grupos de colunas (l) não são conhecidos. Desta forma, foram testados variados números de k e l. Se o número de grupos de um conjunto, seja de linhas ou colunas, for x, os valores utilizados para k e l variaram de 2 até $2x$. A tabela 3 apresenta os valores de k e l utilizados como parâmetros de entrada dos algoritmos, para cada estrutura.

Tabela 3 – Parâmetros k e l utilizados para experimentos com dados sintéticos

Estrutura de Co-grupo	Grupo de linha k	Grupo de coluna l
11a	{2, 3, 4}	{2, 3, 4}
11b	{2, 3, 4, 5, 6}	{2, 3, 4, 5, 6}
11c	{2, 3, 4, 5, 6}	{2, 3, 4, 5, 6}
11d	{2, 3, 4, 5, 6}	{2, 3, 4, 5, 6, 7, 8, 9, 10}
11e	{2, 3, 4, 5, 6, 7, 8, 9, 10}	{2, 3, 4, 5, 6}
11f	{2, 3, 4, 5, 6}	{2, 3, 4, 5, 6}
11g	{2, 3, 4, 5, 6}	{2, 3, 4, 5, 6}
11h	{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}	{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
11i	{2, 3, 4, 5, 6, 7, 8, 9, 10}	{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Para o algoritmo *k-means*, somente o parâmetro k foi utilizado, portanto, apenas a coluna "Grupo de linha k" deve ser considerada. Isto porque o algoritmo não encontra grupos de colunas naturalmente. Para os demais algoritmos, ambos os parâmetros foram utilizados.

Um outro parâmetro de entrada dos algoritmos é o número máximo de iterações. Os valores definidos foram 100, 500, 1.000 e 10.000 iterações. A convergência neste trabalho é alcançada de duas formas: i) quando a diferença do erro de reconstrução da matriz X em duas iterações consecutivas é menor que um limiar, aqui definido como 10^{-4}, ii) quando a condição do item i não é satisfeita e o algoritmo atinge o número máximo de iterações parametrizado.

4.1.3 Estratégia para determinar pertinência a grupos de linhas e colunas

Ao final da execução de um algoritmo, ele gera duas ou três matrizes fatoradas, dependendo da abordagem de fatoração (dupla ou tripla). A composição de algumas dessas matrizes geram bases que representam os grupos de linhas e os grupos de colunas, também chamadas de vetores protótipos. O quadro 2 contém as formulações das bases para cada algoritmo, tanto para os grupos de linhas quanto para os grupos de colunas. O detalhe de cada uma dessas bases está mais bem explicado no capítulo 2, na seção que apresenta cada algoritmo. Essa informação é colocada aqui para apoiar o leitor nas seções que discutem os resultados, pois essa foi a estratégia adotada para determinar a pertinência de uma linha a um grupo de linhas e de uma coluna a um grupo de colunas.

Quadro 2 – Formulação das bases (vetores protótipos) que representam os grupos de linhas e de colunas para cada algoritmo

Algoritmo	Base para grupos de linhas	Base para grupos de colunas
k-means	C	-
NBVD, ONM3F, ONMTF FNMTF, WC-NMTF e WC-FNMTF	SV^T	US
OvNMTF e BinOvNMTF	$\sum_{p=1}^{k} I_{(p)} S V_{(p)}^T$	$U I_{(p)} S$

Fonte: Waldyr Lourenço de Freitas Junior, 2023

4.1.4 Resultados

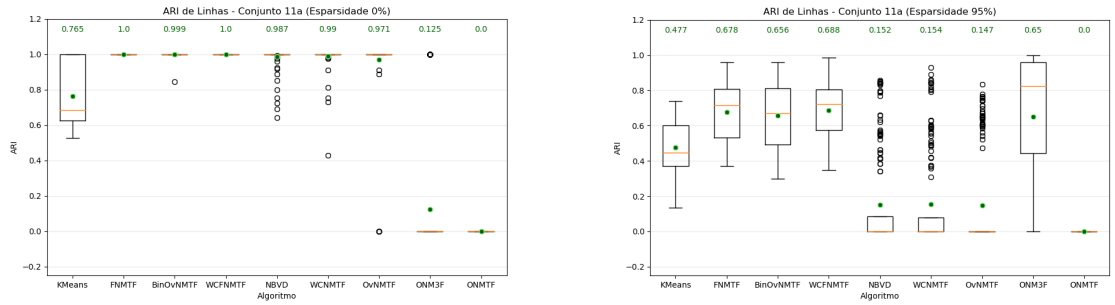
Esta seção tem como objetivo apresentar os resultados dos experimentos realizados com conjuntos de dados sintéticos e desenvolver uma discussão sobre eles. As medidas utilizadas para validação dos agrupamentos foram ARI, tanto de linhas quanto de colunas, índice *Silhouette* e erro de reconstrução gerado pelos algoritmos.

Conforme explanado na seção 4.1.1, foram utilizadas nove estruturas de cogrupos distintas. A estrutura mais simples contém apenas um cogrupo. As estruturas mais complexas possuem diversos cogrupos, sobrepostos ou não. Os gráficos da figura 16 são referentes ao ARI de linhas obtido em execuções dos algoritmos sobre o conjunto 11a (conjunto com a estrutura mais simples). Em ambos os gráficos, são comparados os

resultados de todos os algoritmos. A linha na cor laranja representa a mediana, o círculo preto-esverdeado e o valor numérico acima da caixa representam a média.

A quantidade de pontos do gráfico (número de execuções) varia para cada conjunto de dados de acordo com os parâmetros preestabelecidos: tamanho do conjunto $n \times m$ (100 $\times$ 100, 100 $\times$ 300, 300 $\times$ 100 e 300 $\times$ 300), número de grupo de linhas k (cf. tabela 3), número de grupo de colunas l (cf. tabela 3) e número máximo de iterações (cf. seção 4.1.2). Por exemplo, o gráfico da figura 16a ilustra 144 execuções do NBVD para o conjunto 11a com dados densos (esparsidade 0%): 4 conjuntos $\times$ 3 grupos de linhas $\times$ 3 grupos de colunas $\times$ 4 diferentes parâmetros de iterações.

Figura 16 – Diagrama de caixa para ARI de linhas em experimentos executados sob o conjunto 11a (144 execuções para cada algoritmo em cada gráfico). (a) conjunto 11a com dados densos e (b) conjunto 11a com fator alto de esparsidade



(a) (b)

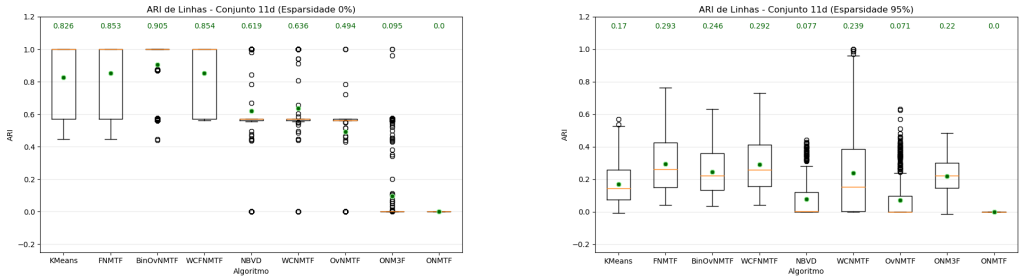
Fonte: Waldyr Lourenço de Freitas Junior, 2023

Os algoritmos FNMTF, BinOvNMTF e WC-FNMTF apresentaram os melhores resultados em ambas as situações (dados densos e dados com fator alto de esparsidade). Houve uma queda de desempenho no caso dos dados com fator de esparsidade maior. Os algoritmos NBVD, WC-NMTF e OvNMTF demonstraram ser muito sensíveis à esparsidade, apresentando uma queda de desempenho alta. Umas das razões para esses resultados pode ser a natureza dos algoritmos. Os primeiros possuem restrições binárias ao passo que os últimos não as possuem. O algoritmo ONM3F, que junto com o ONMTF teve o pior resultado para o caso de dados densos, melhorou seu desempenho para o conjunto com fator alto de esparsidade e alcançou o patamar do FNMTF, do BinOvNMTF e do WC-FNMTF. O algoritmo ONM3F possui restrições de ortogonalidade, e tais restrições favorecem trabalhar com matrizes esparsas, característica de corpus de textos representados como matrizes.

O trabalho de Ding e He (2005), que propôs o algoritmo ONM3F, não discutiu a eficiência dele em alta esparsidade, no entanto, os experimentos deste trabalho de pesquisa demonstraram que à medida que o fator de esparsidade aumentou, no geral, o desempenho do algoritmo também aumentou.

A figura 17 também apresenta dois gráficos de comparação do ARI de linhas, de modo semelhante aos apresentados anteriormente, mas referente ao conjunto 11d. Esse conjunto tem uma particularidade interessante para ser avaliada, que é possuir sobreposição de colunas. Os algoritmos OvNMTF e BinOvNMTF foram concebidos para naturalmente trabalhar com sobreposição de colunas (BRUNIALTI *et al.*, 2017; FREITAS JR. *et al.*, 2020). Apesar do conjunto possuir sobreposição nas colunas, ele não possui sobreposição nas linhas, o que facilita a busca por grupos de linhas pelos algoritmos.

Figura 17 – Diagrama de caixa para ARI de linhas em experimentos executados sob o conjunto 11d (720 execuções para cada algoritmo em cada gráfico). (a) conjunto 11d com dados densos e (b) conjunto 11d com fator alto de esparsidade



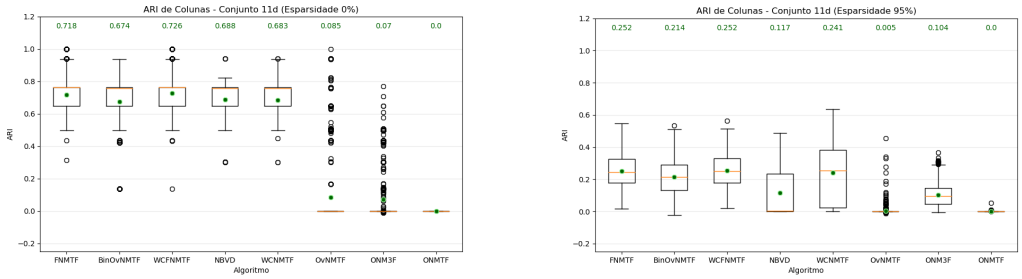
(a) (b)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

O resultado do ARI de linhas para o conjunto 11d teve um comportamento semelhante ao resultado já discutido para o conjunto 11a. O WC-NMTF apresentou algumas execuções, para o conjunto com fator alto de esparsidade, que atingiram o valor máximo para o ARI e seu resultado médio foi próximo dos algoritmos com restrições binárias. Não foram observadas diferenças significativas entre os algoritmos FNMTF e BinOvNMTF; o BinOvNMTF foi levemente melhor para o conjunto com dados densos e o FNMTF foi levemente melhor para o conjunto com fator alto de esparsidade. O NBVD apresentou resultado médio melhor que o resultado do OvNMTF, para o conjunto com dados densos; já para o conjunto com fator alto de esparsidade, o resultado médio foi equivalente, mas o OvNMTF apresentou diversas execuções acima do NBVD.

Devido à particularidade já mencionada e com o objetivo de relacionar os resultados do ARI de linhas e de colunas para o conjunto 11d, a figura 18 apresenta os resultados para o ARI de colunas.

Figura 18 – Diagrama de caixa para ARI de colunas em experimentos executados sob o conjunto 11d (720 execuções para cada algoritmo em cada gráfico). (a) conjunto 11d com dados densos e (b) conjunto 11d com fator alto de esparsidade



(a) (b)

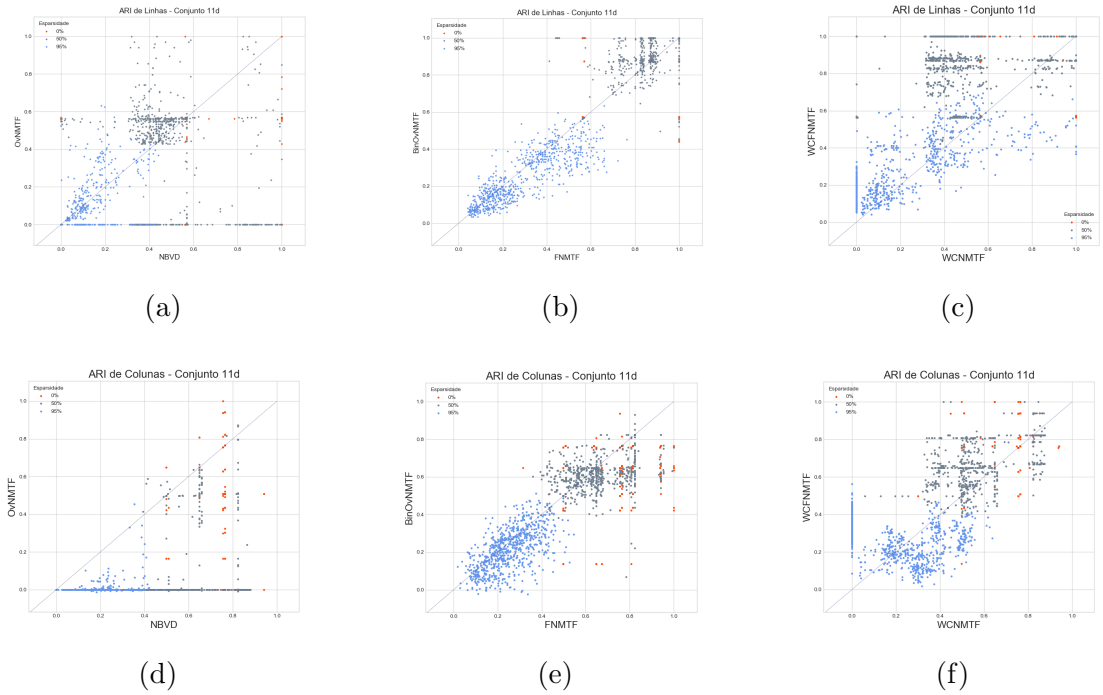
Fonte: Waldyr Lourenço de Freitas Junior, 2023

De uma forma geral, os resultados demonstraram que os algoritmos FNMTF e WC-FNMTF tiveram um desempenho melhor. O WC-NMTF apresentou novamente alguns resultados isolados acima de todos os algoritmos (Conjunto com fator alto de esparsidade), apesar de o resultado médio estar próximo dos resultados do FNMTF e WC-FNMTF. O algoritmo OvNMTF apresentou um resultado médio menor quando comparado ao NBVD, e demonstrou maior instabilidade para ambos os conjuntos. Novamente, alguns resultados isolados foram acima dos resultados do NBVD. A diferença de resultado do BinOvNMTF para o FNMTF foi mais sutil. Esse é um ponto relevante para avaliação dos algoritmos, dado que o OvNMTF é baseado na estratégia de regras de atualização do NBVD, e o BinOvNMTF é baseado na estratégia de regras de atualização do FNMTF, ainda que os trabalhos de Brunialti *et al.* (2017) e Freitas Jr. *et al.* (2020) não usaram o ARI de colunas (para conjunto de dados sintéticos) como medida de avaliação dos algoritmos. Outro ponto relevante dos experimentos foi o resultado do algoritmo aqui proposto, o WC-FNMTF, que teve o melhor resultado diante dos demais, inclusive sobre o seu algoritmo base (WC-NMTF), que demonstrou ser mais instável.

Para explorar um pouco mais o ponto destacado no parágrafo anterior, a figura 19 apresenta uma visão mais detalhada da comparação do desempenho do NBVD com OvNMTF (figuras 19a e 19d), do FNMTF com BinOvNMTF (figura 19b e 19e) e do WC-NMTF com o WC-FNMTF (figura 19c e 19f), para ARI de linhas e colunas para

o conjunto 11d. O eixo x do gráfico representa o ARI de um dos algoritmos e o eixo y representa o ARI do outro algoritmo envolvido na comparação. A linha reta $x = y$ é uma linha divisória. Os pontos abaixo da linha significam que o algoritmo do eixo x apresentou um desempenho melhor. Os pontos acima da linha significam o oposto. O ARI varia de -1 a 1, contudo, para ter uma visão mais detalhada dos resultados, considerando que não houve resultados abaixo de zero, a escala do gráfico foi alterada para demonstrar esse detalhe.

Figura 19 – Comparativo do ARI de linhas, conjunto 11d, para os algoritmos (a) NBVD e OvNMTF, (b) FNMTF e BinOvNMTF, (c) WC-NMTF e WC-FNMTF, e ARI de colunas para os algoritmos (d) NBVD e OvNMTF, (e) FNMTF e BinOvNMTF, (f) WC-NMTF e WC-FNMTF.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

Estes gráficos apresentam os resultados de todas as execuções para os conjuntos com dados densos e com fatores de esparsidade médio e alto. O gráfico da figura 19a, que apresenta os resultados para o ARI de linhas, visualmente demonstra que o algoritmo OvNMTF alcançou melhores resultados que o NBVD, especialmente para esparsidade maior, apesar de que existem diversos resultados zerados para OvNMTF. Já a figura 19d demonstra claramente um desempenho melhor para NBVD e é possível notar que apenas uma execução do OvNMTF foi superior a do NBVD.

Os gráficos das figuras 19b e 19e demonstram desempenho sutilmente melhor para FNMTF. Os gráficos das figuras 19c e 19f também demonstram um desempenho sutilmente melhor para WC-FNMTF, mas no caso de ARI de colunas, um pouco mais evidente nos conjuntos com baixa esparsidade. Essas observações corroboram a discussão da figura 18. Os autores que propuseram os algoritmos OvNMTF e BinOvNMTF não avaliaram os resultados por meio do ARI de colunas.

A figura 20 apresenta outros gráficos de comparativo de desempenho dos algoritmos. As figuras 20a, 20b e 20c demonstram o desempenho de *k-means*, FNMTF e BinOvNMTF sob a ótica do ARI de linhas. A decisão de comparar esses três algoritmos se dá por conta da natureza deles, que é de restrição binária (para pertinência a grupos). Além disso, a estratégia de implementação das regras de atualização, para os três algoritmos, é similar. O BinOvNMTF se baseia na estratégia do FNMTF, e este último se baseia na estratégia de *k-means*. O conjunto 11g tem uma particularidade interessante de sobreposição de linhas e sobreposição de colunas, simultaneamente. Tanto o FNMTF quanto o BinOvNMTF foram superiores ao *k-means*. A principal diferença entre eles é que este último é baseado em uma abordagem de fatoração dupla de matrizes, e algoritmos que usam abordagem de fatoração tripla aparentemente resolvem os problemas mais facilmente, levando em consideração o maior detalhamento proporcionado pelas matrizes fatoradas. Não há diferenças significativas[1] entre o desempenho do FNMTF e do BinOvNMTF.

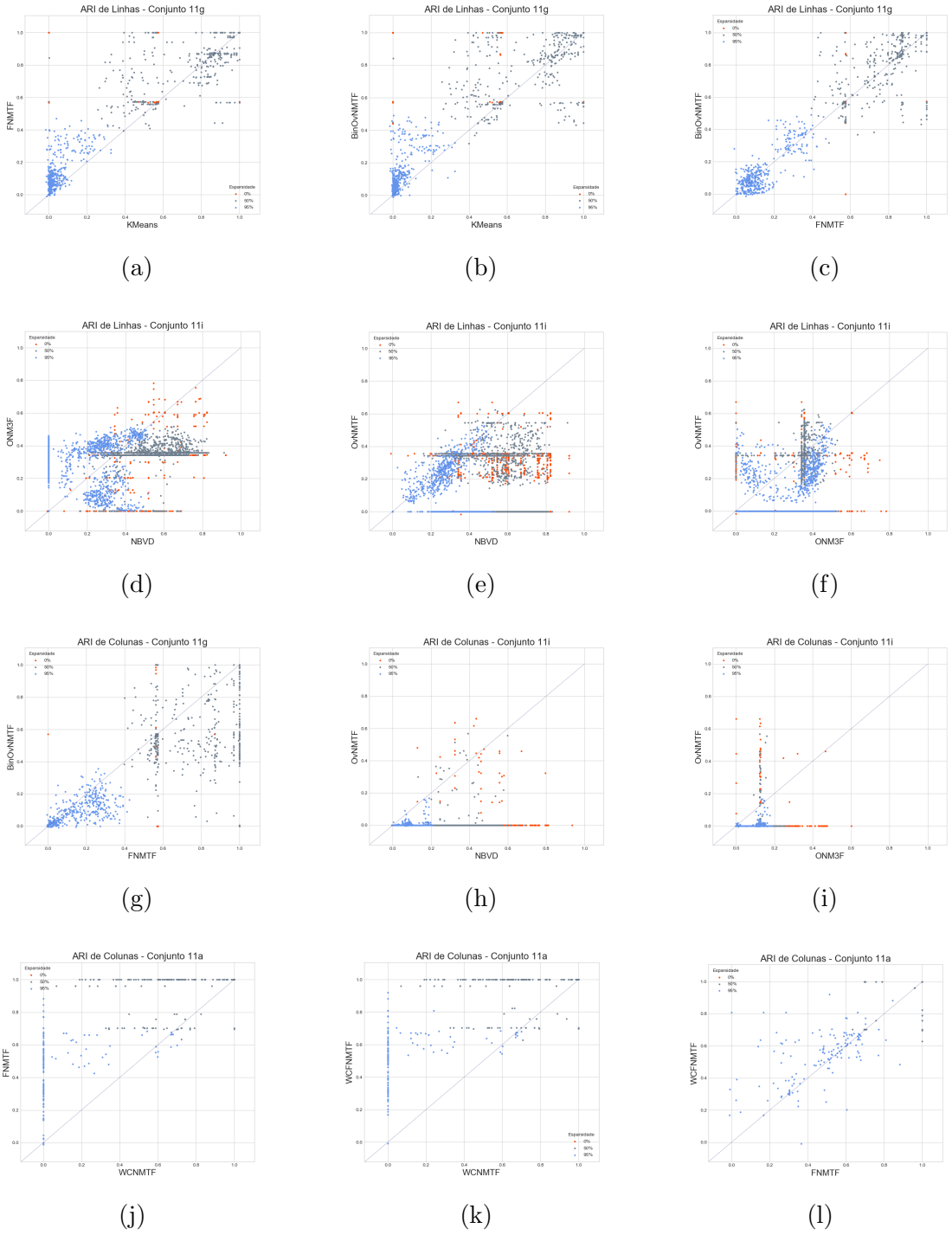
As figuras 20d, 20e e 20f demonstram o desempenho de NBVD, ONM3F e OvNMTF para o ARI de linhas. A razão de escolhê-los também se dá por conta da natureza deles, a ideia foi comparar algoritmos da mesma "família". Foi escolhido o conjunto 11i para esse comparativo, que é um conjunto com sobreposição de linhas, sobreposição de colunas e com cogrupos sobrepostos.

De uma forma geral, o algoritmo NBVD demonstrou ter um desempenho melhor perante ONM3F e OvNMTF, entretanto, à medida que a esparsidade aumentou, o desempenho do NBVD se equalizou com os outros dois. Isso demonstra que o NBVD pode ser mais sensível à alta esparsidade. O trabalho de Freitas Jr. *et al.* (2020) também relata aspectos positivos do OvNMTF com relação à alta esparsidade no conjunto de dados, mas importante destacar que somente o ARI de linhas foi avaliado no estudo citado.

Para fazer um paralelo com a figura 20c, a figura 20g compara o desempenho do FNMTF com BinOvNMTF sob a ótica do ARI de colunas. Não é possível realizar esse

[1] Do ponto de vista de interpretação simples. Não foram executados testes estatísticos de significância.

Figura 20 – Gráficos comparativos entre dois algoritmos para ARI de linhas e ARI de colunas em experimentos executados sob os conjuntos 11g e 11i.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

paralelo considerando *k-means*, pois este último não gera grupos de colunas e não tem resultados de ARI de colunas. Nesses experimentos, o FNMTF apresentou um desempenho melhor que o BinOvNMTF, comportamento que também requer maior investigação, pois

como já explanado, os autores do BinOvNMTF (BRUNIALTI *et al.*, 2017) não realizaram experimentos validando ARI de colunas.

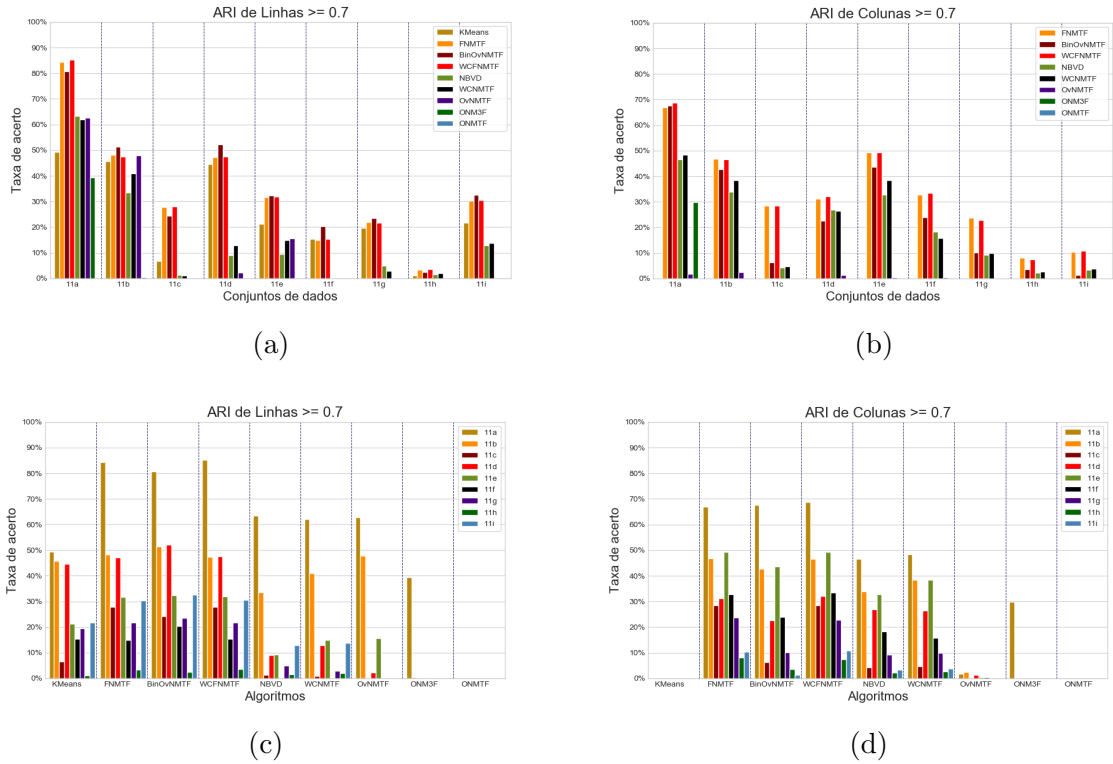
O mesmo paralelo foi feito entre as figuras 20e e 20h, e entre as figuras 20f e 20i. O algoritmo NBVD demonstrou novamente, no geral, um desempenho melhor comparado ao OvNMTF, dessa vez para ARI de colunas. Inclusive, dessa vez o OvNMTF não conseguiu ter resultados equivalentes para alta esparsidade nem contra o NBVD, nem contra o ONM3F. As particularidades do conjunto 11i podem ter influenciado tais resultados, visto que tais algoritmos ainda não tinham sido submetidos a experimentos em conjuntos com sobreposição de linhas, de colunas e com cogrupos sobrepostos.

As figuras 20j, 20k e 20l demonstram o desempenho de FNMTF, WC-NMTF e WC-FNMTF para o ARI de colunas. O conjunto utilizado é o conjunto 11a, que é o conjunto que possui um único cogrupo. Os algoritmos FNMTF e WC-FNMTF apresentaram um desempenho melhor do que o desempenho do algoritmo WC-NMTF. O resultado deles (FNMTF e WC-FNMTF) foram muito similares entre si, entretanto, em alta esparsidade, levemente melhor para o algoritmo que foi concebido neste trabalho.

Uma visão mais generalista dos resultados de ARI de linhas (figuras 21a e 21c) e colunas (figuras 21b e 21d) pode ser observada na figura 21. Somente resultados de ARI de linhas ou colunas maiores que 0,70 são considerados. Os gráficos apresentam duas visões diferentes para os mesmos resultados, uma por conjunto de dados e outra por algoritmo. Nas figuras 21a e 21b, cada setor do gráfico representa um conjunto de dados. Dentro de cada setor são apresentados os resultados de cada algoritmo. O eixo y representa a taxa de acerto de determinado algoritmo. Essa taxa é medida pelo número de execuções nas quais o algoritmo atingiu o ARI maior ou igual a 0,70, dividido pelo número total de execuções. O algoritmo *k-means* não é considerado nos resultados das figuras 21b e 21d. O mesmo raciocínio pode ser aplicado para as figuras 21c e 21d, em que cada setor do gráfico representa um algoritmo e dentro de cada setor são apresentados os resultados por conjunto.

De maneira geral, os algoritmos com restrições binárias tem um desempenho melhor tanto para ARI de linhas quanto para ARI de colunas. O BinOvNMTF teve o melhor desempenho para ARI de linhas, alcançando a maior taxa de acerto em 6 conjuntos. Corroborando com as análises previamente apresentadas, o desempenho dos algoritmos FNMTF e WC-FNMTF ficou muito similar. Além das restrições binárias dos algoritmos,

Figura 21 – Comparativo da taxa de acerto dos algoritmos para ARI maiores que 0,70. (a) ARI de linhas por conjunto de dados, (b) ARI de colunas por conjunto de dados, (c) ARI de linhas por algoritmo e (d) ARI de colunas por algoritmo.



Fonte: Waldyr Lourenço de Freitas Junior, 2023

o BinOvNMTF tem como diferencial as k matrizes V, que aumenta a flexibilidade do agrupamento; essa característica influenciou o resultado do algoritmo para o ARI de linhas.

O melhor desempenho para o ARI de colunas também foi dos algoritmos com restrições binárias. O WC-FNMTF atingiu o melhor desempenho em mais da metade dos conjuntos, seguido do FNMTF; o fator extra (matriz de coocorrência) do algoritmo no processo de agrupamento permite organizar melhor os grupos de colunas. O BinOvNMTF foi melhor que o FNMTF apenas no conjunto 11a. Os algoritmos NBVD e WC-NMTF atingiram melhores resultados para ARI de colunas do que para ARI de linhas, já o comportamento do OvNMTF foi o contrário.

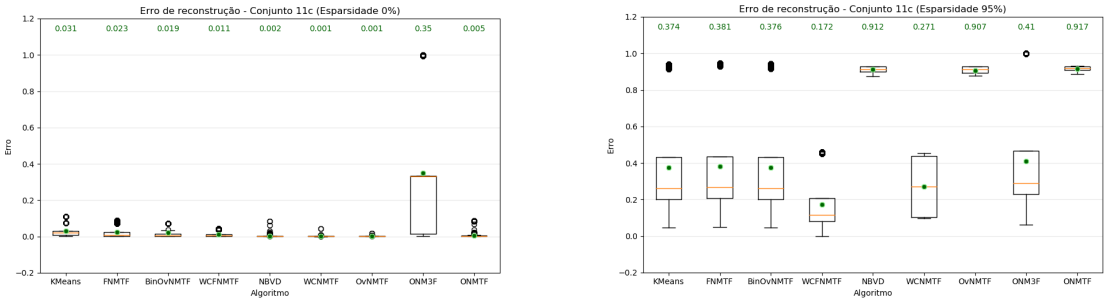
Um comportamento interessante observado neste gráfico é que os resultados do conjunto 11d para o ARI de linhas foram melhores que os resultados para o ARI de colunas, e os resultados do conjunto 11e tiveram o comportamento inverso, melhores para o ARI de colunas. A diferença principal de ambos os conjuntos é que o conjunto 11d possui sobreposição nas colunas e as linhas são exclusivas, e o conjunto 11e possui sobreposição

nas linhas e as colunas são exclusivas. Em outras palavras, a sobreposição de linhas ou colunas em um conjunto possui um efeito direto no desempenho do algoritmo.

Especificamente para o ARI de linhas, esse efeito deveria ser minimizado pelo algoritmo OvNMTF, que como já explanado, foi criado para tratar sobreposição de colunas, a fim de permitir mais flexibilidade no agrupamento das linhas. Entretanto, os resultados demonstraram que o NBVD alcançou melhores resultados no conjunto com essa característica (11d).

Um outro ponto importante para análise de um algoritmo é a capacidade que ele tem de reconstruir a matriz original, após o processo de fatoração. Essa capacidade é avaliada pelo erro de reconstrução, que é calculado pela diferença entre a matriz original e a matriz reconstruída. A figura 22 apresenta o comportamento do erro de reconstrução para o conjunto 11c, um conjunto com nove cogrupos de interesse. O erro foi normalizado usando a função *MinMaxScaler()* do Python. Essa função normaliza o erro para um valor entre 0 e 1, sendo 0 (zero) o melhor caso (menor erro) e 1 (um) o pior caso (maio erro), por meio do seguinte cálculo: $(x - min)/(max - min)$, em que x é o valor do erro a ser normalizado, min é o erro mínimo da distribuição e max é o erro máximo.

Figura 22 – Diagrama de caixa para erro de reconstrução em experimentos executados sob o conjunto 11c (400 execuções para cada algoritmo em cada gráfico) para (a) conjunto 11c com dados densos e (b) conjunto 11c com fator alto de esparsidade.



(a) (b)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

O algoritmo WC-FNMTF, proposto neste trabalho, atingiu o menor erro de reconstrução para o conjunto com fator alto de esparsidade. Esse comportamento também foi observado em experimentos com outros conjuntos. Outra observação na figura 22 é que para o conjunto com fator alto de esparsidade, os algoritmos com restrições binárias apresentaram resultados médios menores. Isso indica que ao reconstruir a matriz original,

tais algoritmos o fazem com mais precisão que os demais. O algoritmo WC-NMTF proposto no trabalho de Salah, Ailem e Nadif (2018) apresenta o erro mais baixo entre os algoritmos sem restrições binárias, seguido do algoritmo ONM3F.

Outro comportamento observado foi que o erro de reconstrução para os algoritmos OvNMTF e BinOvNMTF foi sutilmente melhor que os algoritmos NBVD e FNMTF, respectivamente. Os primeiros são baseados nas estratégias de regras de atualização dos últimos, respectivamente.

4.1.5 Considerações Finais

Os algoritmos que possuem restrições binárias em sua arquitetura apresentaram resultados mais satisfatórios do que os que não as possuem, para o ARI de linhas, ARI de coluna e erro de reconstrução. FNMTF e BinOvNMTF mostraram-se equivalentes na maioria dos resultados. WC-FNMTF mostrou-se superior na maior parte dos experimentos, sobretudo na reconstrução da matriz original em esparsidade alta. Dentre essa família de algoritmos, *k-means* foi o que ficou com o pior desempenho.

Dentre os algoritmos que não possuem as restrições binárias, os que tiveram melhor desempenho foram WC-NMTF e ONM3F, este último para conjuntos com alta fator de esparsidade. Os demais não apresentaram resultados significativos.

O quadro 3 apresenta o resultado consolidado dos experimentos realizados com os dados sintéticos. O "x" representa que o algoritmo teve um bom desempenho nos experimentos, considerando a medida de avaliação relacionada, ou seja, para o ARI de linhas, ARI de colunas e erro de reconstrução.

Quadro 3 – Quadro consolidado dos experimentos com dados sintéticos

Algoritmo	Dados densos			Dados esparsos		
	ARI linhas	ARI colunas	Erro	ARI linhas	ARI colunas	Erro
K-means	x					
NBVD		x				
ONM3F				x	x	
ONMTF						
OvNMTF			x			
FNMTF	x	x		x	x	
BinOvNMTF	x	x		x	x	
WC-NMTF		x	x	x	x	
WC-FNMTF	x	x	x	x	x	x

Fonte: Waldyr Lourenço de Freitas Junior, 2023

4.2 Experimento com dados do mundo real

Esta seção apresenta os experimentos realizados com dados do mundo real. A seção 4.2.1 apresenta o corpus utilizado. A seção 4.2.2 discute o pré-processamento realizado sobre ele. Na seção 4.2.3 é apresentada a configuração dos experimentos. A seção 4.2.4 apresenta uma discussão dos resultados decorrentes dos experimentos. Por fim, a seção 4.2.5 faz as considerações finais.

4.2.1 Conjunto de dados - hiperpartidarismo

O corpus utilizado neste trabalho é uma coletânea de notícias extraídas de diferentes portais de notícias, majoritariamente entre os anos de 2016 e 2018. Os dados foram publicados como parte de uma tarefa do SemEval (*International Workshop on Semantic Evaluation*)[2] de 2019 e podem ser acessados pela plataforma Zenodo[3]. Os dados estão rotulados em textos hiper partidários e não hiper partidários, de acordo com a definição de hiperpartidarismo dos próprios autores. Tal conjunto foi rotulado por meio de *crowdsourcing*. Ele foi gerado somente com os textos para os quais houve consenso de rotulação entre os voluntários do *crowdsourcing*.

O corpus é composto por 645 notícias, das quais 238 (36,9%) estão rotuladas como hiper partidárias e 407 (63,1%) estão rotuladas como não hiper partidárias, como apresentado na tabela 4.

Tabela 4 – Organização do corpus de notícias hiper partidárias

Ano	Hiper partidário	Não hiper partidário
2018	18	40
2017	159	224
2016	55	97
< 2016	2	19
Sem data	4	27

Fonte: Waldyr Lourenço de Freitas Junior, 2023

O conjunto de dados está no formato XML, com os principais dados das notícias: identificador, data da publicação, título e o texto da notícia em si. A figura 23 mostra um exemplo de uma notícia nesse formato e a figura 24 mostra a notícia publicada pelo site

[2] https://semeval.github.io/
[3] https://zenodo.org/record/1489920#.Y06AsnbMI2z

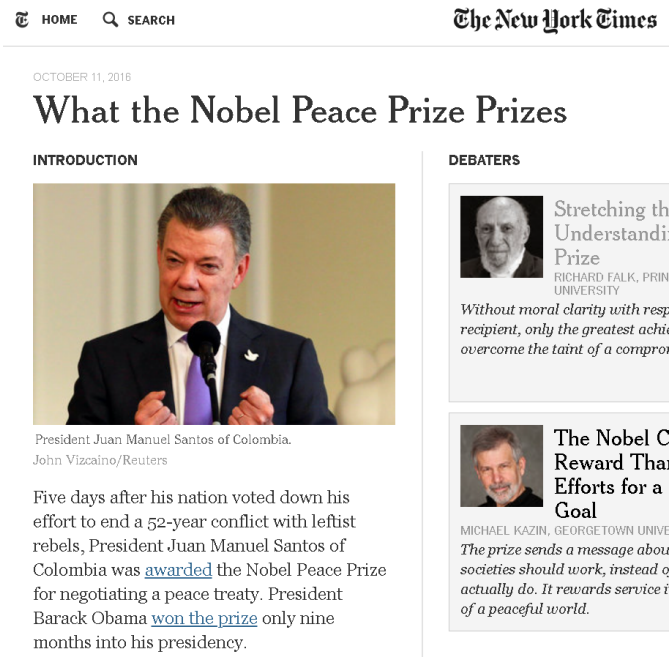
The New York Times[4]. Um outro arquivo em formato XML contendo o rótulo da notícia e seu respectivo link também foi disponibilizado pelos autores do conjunto.

Figura 23 – Exemplo de notícia em formato XML para o conjunto de dados do mundo real

```
<article
    id = "0000365"
    published-at = "2016-10-11"
    title = "What the Nobel Peace Prize Prizes">
        President Juan Manuel Santos of Colombia.
        <p>John Vizcaino/Reuters</p>
        <p>Five days after his nation voted down his
        effort to end a 52-year conflict with leftist
        rebels, President Juan Manuel Santos of
        Colombia was <a type="internal">awarded</a> the Nobel Peace Prize
        for negotiating a peace treaty. President
        Barack Obama <a type="internal">won the prize</a> only nine
        months into his presidency.</p>
```

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Figura 24 – Recorte da notícia exemplificada na figura 23, extraída direto do site *The New York Times*



Fonte: The New York Times, 2016

O conjunto é originalmente não balanceado e possui 20.817 palavras distintas, das quais 16.677 com frequência menor e igual a sete. Neste trabalho, essas palavras com baixa frequência foram classificadas como *stopwords* (cf. seção 4.2.2).

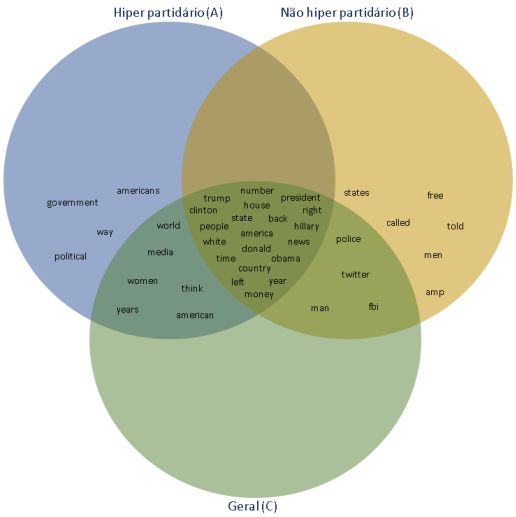
Para os experimentos deste trabalho, foi gerado um corpus balanceado para os testes, baseado no corpus original. Foram selecionadas aleatoriamente 238 notícias não hiper partidárias para que o conjunto ficasse balanceado. Esse novo conjunto possui 18.348 palavras distintas, das quais 14.460 possuem frequência menor e igual a seis. Tais palavras de baixa frequência também foram classificadas como *stopwords*. A escolha das

[4] https://www.nytimes.com/roomfordebate/2016/10/11/what-the-nobel-peace-prize-prizes

stopwords com baixa frequência considerou o resultado final do corpus, objetivando manter a dimensão final (quantidade de palavras distintas) dele entre 3.800 e 4.200 palavras.

Além disso, foram selecionadas as 15 e 30 palavras mais frequentes das notícias hiper partidárias, das notícias não hiper partidárias e do corpus como um todo. Tais palavras foram colocadas em um diagrama de Venn para avaliar a interseção da ocorrência delas nos três subconjuntos. A figura 25 exemplifica essa análise para o conjunto não balanceado com as *top* 30 palavras.

Figura 25 – Diagrama de Venn para as *top* 30 palavras mais frequentes do conjunto não balanceado



Fonte: Waldyr Lourenço de Freitas Junior, 2023

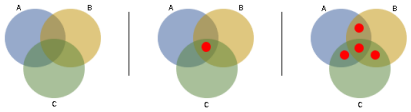
Baseado nesta análise, foram gerados alguns conjuntos para serem utilizados nos experimentos, objetivando maior variabilidade.

- Conjunto com todas as palavras, ou seja, sem corte algum de palavras (X)
- Conjunto com corte das palavras mais frequentes em $A \cap B \cap C$ (Y)
- Conjunto com corte das palavras mais frequentes em $A \cap B$, $A \cap C$, $B \cap C$ e $A \cap B \cap C$ (Z)

A figura 26 demonstra os conjuntos supracitados e a dimensão para cada um deles. Por exemplo, o conjunto balanceado Y para as *top* 15 palavras possui ao todo 476 notícias (238 hiper partidárias e 238 não hiper partidárias) e dimensão 3.878 (palavras).

Com relação à representação vetorial do corpus, foram utilizadas as representações binária, TF e TF-IDF, gerando conjuntos distintos para entrada dos algoritmos. Ao todo, foram gerados 30 conjuntos distintos para os experimentos.

Figura 26 – Relação das dimensões dos conjuntos gerados para os experimentos



		Corte (●) das palavras mais frequentes		
		X	Y	Z
Balanceado	Top 15	3.888	3.878	3.873
	Top 30		3.870	3.858
Não Balanceado	Top 15	4.140	4.129	4.125
	Top 30		4.120	4.110

Fonte: Waldyr Lourenço de Freitas Junior, 2023

4.2.2 Pré-processamento

A etapa de pré-processamento é uma das etapas fundamentais na formatação e representação de textos. Segundo o estudo de Manning, Raghavan e Schütze (2008), algumas das tarefas consideradas nas etapas de pré-processamento são: i) tokenização, ii) remoção de *stopwords*, iii) normalização, iv) redução para o radical[5] e v) lematização.

Para a realização dos experimentos preliminares, foram aplicados os procedimentos descritos a seguir.

Preparação dos dados

Os textos das notícias passaram por uma etapa inicial de limpeza e organização dos dados, objetivando otimizar o pré-processamento.

Os caracteres de quebra de linha, as *tags* de HTML (parágrafos, citações, hiperlinks internos e externos), as acentuações e as pontuações foram retirados do texto. Os valores numéricos foram convertidos para a palavra “number”, valores monetários para a palavra “money” e hora para a palavra “hour”. Por fim, todos os caracteres foram reduzidos para minúsculo.

[5] Para este trabalho foi uma escolha não reduzir as palavras para seu radical. Como um dos objetivos do trabalho é avaliar os resultados dos algoritmos sob uma ótica de interpretabilidade, manter a palavra em seu formato original foi necessário para que fosse possível prover mecanismos de interpretação do conteúdo dos textos organizados nos grupos.

Tokenização

Dada uma sequência de caracteres, o objetivo é dividir essa sequência em pedaços, que são chamados de *tokens*. No contexto deste trabalho, um *token* é definido como uma palavra e o espaço em branco foi definido como caractere divisor das palavras.

Remoção de stopwords

Em processamento de textos existem palavras com frequência muito elevada. Algumas dessas palavras são irrelevantes para o resultado do processamento e são eliminadas na etapa de pré-processamento. Essas palavras são conhecidas na literatura como *stopwords*.

Neste trabalho, foi utilizada uma lista padrão de *stopwords* da biblioteca NLTK[6] do Python, no idioma inglês. Além disso, algumas palavras foram incluídas nessa lista:

- Palavras que foram tratadas na etapa de preparação dos textos: *number*, *hour* e *money*.
- Palavras que possuem frequência muito baixa ou muito alta no texto, e sua retirada não causa prejuízo de qualidade dos dados (estratégia detalhada na seção 4.2.1).

4.2.3 Configuração dos experimentos

Os mesmos algoritmos utilizados para os experimentos com os dados sintéticos foram utilizados para os experimentos com dados do mundo real, a saber: *k-means*, NBVD, ONM3F, ONMTF, OvNMTF, FNMTF, BinOvNMTF, WC-NMTF e WC-FNMTF. Conforme detalhado na seção 4.2.1, o corpus possui dois rótulos (hiper partidário e não hiper partidário), entretanto, além do número de grupos de linhas (k) igual a 2, foram definidos também os valores 3 e 4 para os experimentos. Como não existe nenhuma evidência do número de grupos de colunas (l) existente neste corpus, foram utilizados valores distintos para l, variando de 2 até 4. Portanto, para cada representação vetorial, foram utilizados os valores de k e l iguais a 2, 3 e 4 (Apenas o parâmetro k foi utilizado para o algoritmo *k-means*).

[6] https://www.nltk.org/

O número máximo de iterações dos algoritmos é um parâmetro do experimento. Os valores definidos foram 100 e 1.000 iterações. A convergência neste experimento é igual à que foi definida nos experimentos com dados sintéticos. Ela é alcançada quando a diferença do erro de reconstrução em duas iterações consecutivas é menor que 10^{-4} ou quando essa condição não é satisfeita e o algoritmo atinge o número máximo de iterações. A mesma estratégia para determinar pertinência a grupos, explicada na seção 4.1.3, foi utilizada para este corpus de notícias hiper partidárias.

4.2.4 Resultados

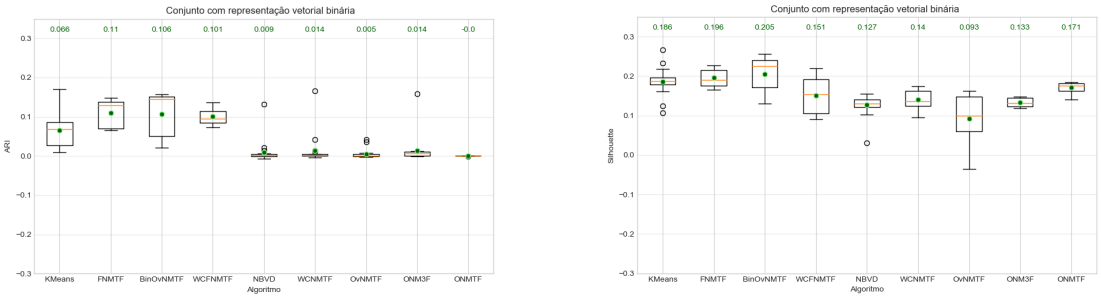
Esta seção tem como objetivo apresentar os resultados dos experimentos realizados com o conjunto de dados de notícias hiper partidárias, previamente apresentado. Os resultados dos experimentos aqui discutidos são feitos sob a ótica das medidas ARI de linhas, índice *Silhouette* e erro de reconstrução. A razão de não usar o ARI de colunas nestes experimentos é a falta de rótulos relacionados a grupo de palavras. Até onde foi possível averiguar para este trabalho de pesquisa, inexistem conjuntos de dados textuais do mundo real com esse tipo de rótulo.

Semelhante ao que foi previamente explicado para os conjuntos de dados sintéticos, na seção 4.2.3, a quantidade de pontos do gráfico (número de execuções) varia de acordo com os parâmetros preestabelecidos. A figura 27 apresenta os resultados dos algoritmos para o conjunto de dados com representação vetorial binária. A figura 27a apresenta os resultados para o ARI de linhas e a figura 27b apresenta os resultados para o índice *Silhouette*.

Os algoritmos com restrições binárias (*k-means*, FNMTF, BinOvNMTF e WC-FNMTF) se destacaram nos resultados do ARI de linhas diante dos demais, apesar dessa diferença ser pequena. As restrições dos algoritmos e a representação vetorial podem ter favorecido este resultado. Nota-se algumas execuções casuais, para os algoritmos sem restrições binárias, com resultados próximos aos alcançados pelos primeiros. Para o índice *Silhouette*, os resultados dos algoritmos com restrições binárias foram muito próximos dos resultados dos algoritmos sem restrições, sem nenhum destaque especial.

A figura 28 consolida os resultados gráficos para os conjuntos com as representações vetoriais TF e TF-IDF. As figuras 28a e 28b apresentam os resultados dos experimentos

Figura 27 – Diagrama de caixa para (a) ARI de linhas e (b) índice *Silhouette* em experimentos realizados sob o conjunto de notícias hiper partidárias com representação vetorial binária.

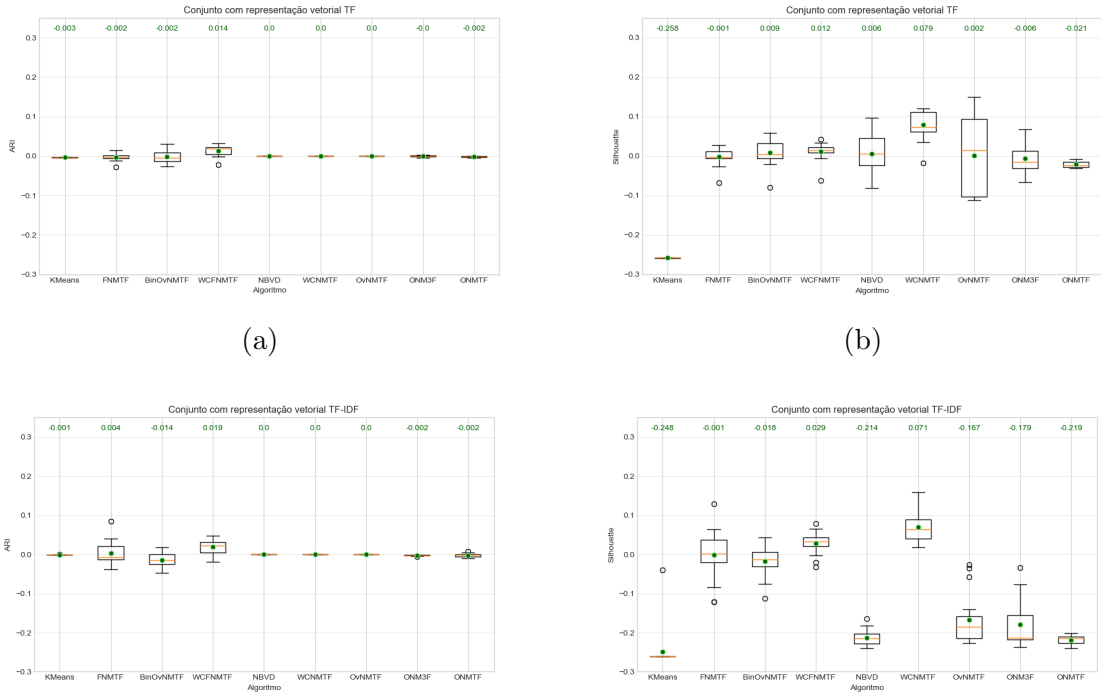


(a) (b)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

com a representação TF para o ARI de linhas e índice *Silhouette*, respectivamente. Já as figuras 28c e 28d, o fazem para a representação TF-IDF.

Figura 28 – Diagrama de caixa para (a) ARI de linhas e (b) índice *Silhouette* com representação TF, e (c) ARI de linhas e (d) índice *Silhouette* com representação TF-IDF, em experimentos realizados sob o conjunto de notícias hiper partidárias.



(a) (b)

(c) (d)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

A respeito da análise da representação TF, os resultados dos algoritmos para o ARI ficaram muito próximos de zero. Os algoritmos BinOvNMTF e WC-FNMTF apresentaram

algumas execuções *outliers*, mas o resultado médio ainda ficou próximo dos demais. O WC-FNMTF foi o único algoritmo com ARI médio positivo. Já os resultados do índice *Silhouette* apresentaram maior variabilidade. Os resultados dos algoritmos com restrições binárias demonstraram menor dispersão e resultados mais próximos de zero. Os resultados dos algoritmos que não possuem restrições binárias variaram mais, com exceção do algoritmo ONMTF. O algoritmo WC-NMTF apresentou o melhor resultado médio, o único com apenas uma execução abaixo de zero.

Já o comportamento observado para os resultados sob o corpus com a representação TF-IDF (figuras 28c e 28d) foi muito semelhante ao comportamento observado anteriormente. Para o ARI, os algoritmos de coagrupamento com restrições binárias tiveram melhores resultados, dos quais o WC-FNMTF alcançou o melhor resultado. Para o índice *Silhouette*, no geral, houve novamente maior variabilidade. O algoritmo WC-NMTF teve o melhor resultado mais uma vez.

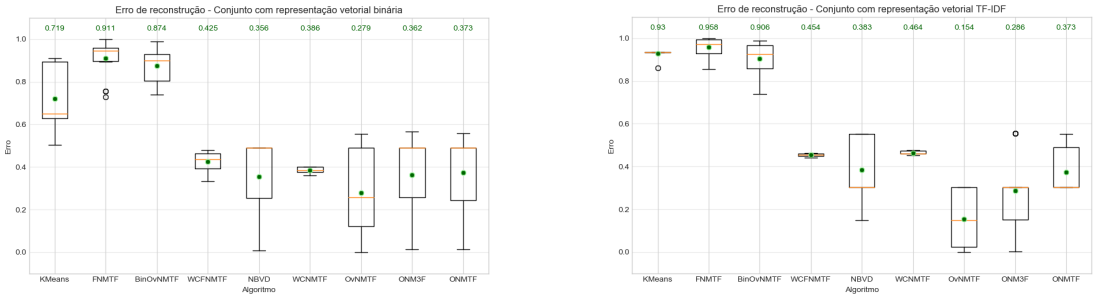
Do ponto de vista de erro de reconstrução, os algoritmos sem restrições binárias demonstraram uma capacidade maior em reconstruir a matriz original. Os resultados foram normalizados usando a função *MinMaxScaler()*[7] do Python, que normaliza o erro para um valor entre 0 e 1, sendo 0 (zero) o melhor caso (menor erro) e 1 (um) o pior caso (maio erro). A figura 29a e a figura 29b demonstram os resultados dos experimentos. Os algoritmos *k-means*, FNMTF e BinOvNMTF tiveram os piores resultados para ambos os conjuntos, com representação vetorial binária e TD-IDF. Os algoritmos WC-NMTF e WC-FNMTF demonstraram baixa variação, mas ainda com baixa capacidade de reconstrução da matriz original.

A figura 30 apresenta a dispersão dos resultados do ARI de linhas e índice *Silhouette* para os algoritmos BinOvNMTF e FNMTF. As figuras 30a, 30b e 30c comparam ambos os algoritmos para o ARI. As figuras 30d, 30e e 30f o fazem para o índice *Silhouette*. A escala do gráfico foi alterada objetivando dar uma visão mais detalhada dos resultados. Os resultados de ambos os algoritmos estão muito próximos, sem quaisquer destaques para um ou outro.

Na figura 31 é apresentada a comparação entre o algoritmo WC-NMTF, proposto no trabalho de Salah, Ailem e Nadif (2018), e WC-FNMTF, proposto neste trabalho. As

[7] $(x - min)/(max - min)$, em que x é o valor do erro a ser normalizado, min é o erro mínimo da distribuição e max é o erro máximo

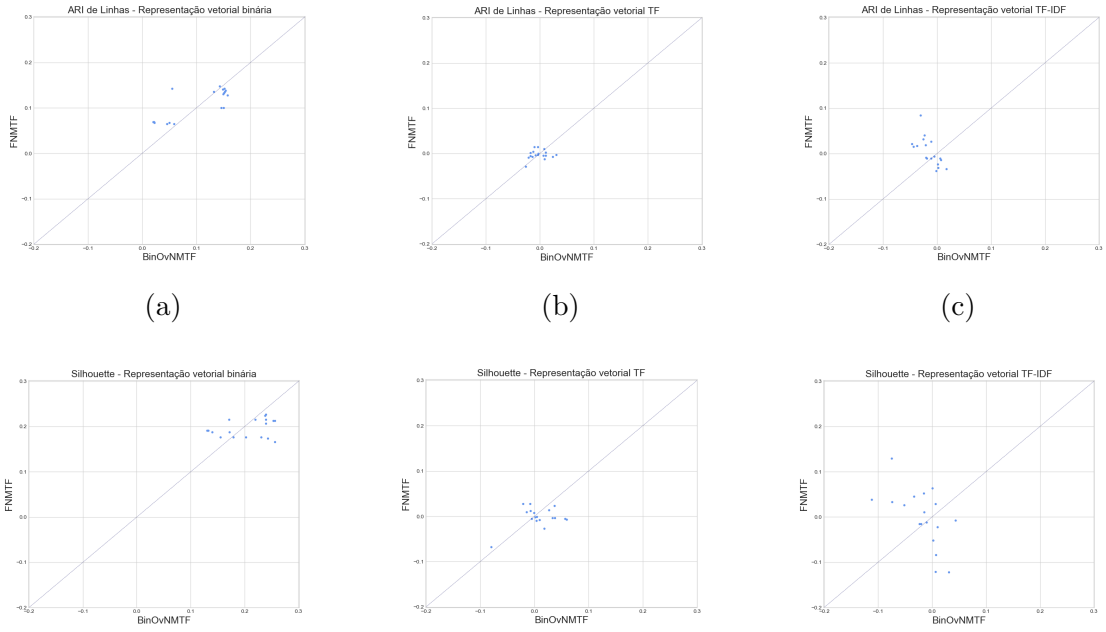
Figura 29 – Diagrama de caixa para erro de reconstrução em experimentos realizados sob o conjunto de notícias hiper partidárias para (a) conjunto com representação vetoriais binária e (b) conjunto com representação vetorial TF-IDF.



(a) (b)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Figura 30 – Gráficos para comparar os algoritmos FNMTF e BinOvNMTF por meio do ARI de linhas (gráficos (a), (b) e (c)) e do índice *Silhouette* (gráficos (d), (e) e (f)), em experimentos executados sob o conjunto de notícias hiper partidárias com as representações vetoriais binária, TF e TF-IDF



(a) (b) (c)

(d) (e) (f)

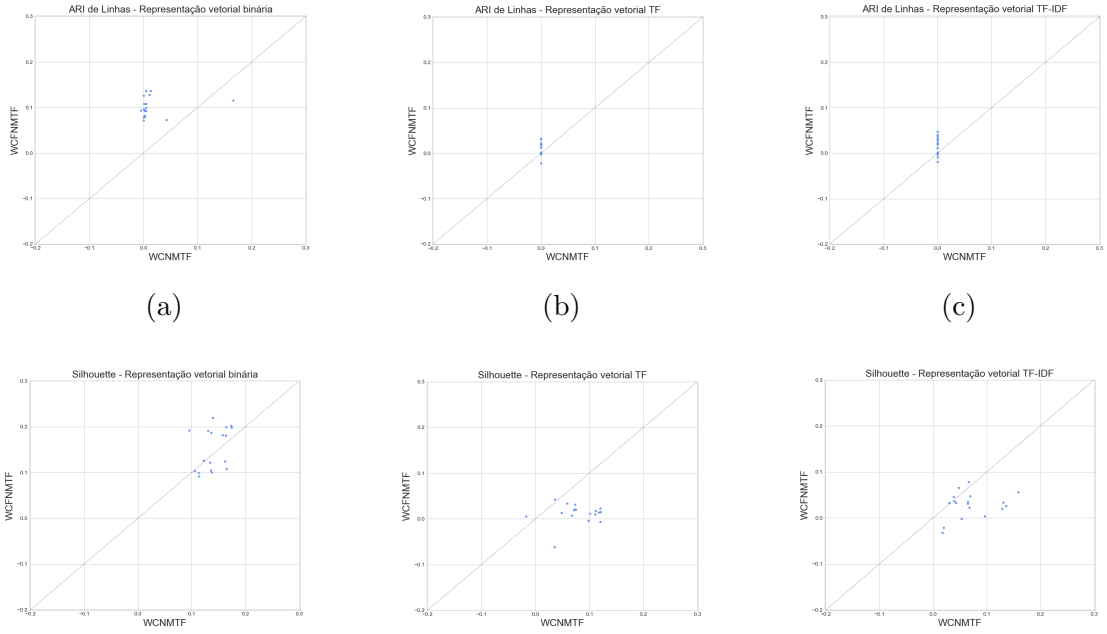
Fonte: Waldyr Lourenço de Freitas Junior, 2023

três primeiras figuras (31a, 31b e 31c) comparam o resultado do ARI de linhas e as últimas três (31d, 31e e 31f) comparam o resultado do índice *Silhouette*.

O algoritmo WC-FNMTF teve resultados superiores para o ARI de linhas, sobretudo para a representação binária, em que apenas uma execução do WC-NMTF foi melhor. O resultado do algoritmo WC-NMTF foi superior para o índice *Silhouette*, especialmente

para os conjuntos com representações baseadas em frequência. A representação binária apresentou um resultado equilibrado. Pela leitura dos resultados, pode-se inferir que o algoritmo WC-FNMTF, que possui restrições binárias, teve um comportamento melhor para o conjunto com representação vetorial binária.

Figura 31 – Gráficos para comparar os algoritmos WC-NMTF e WC-FNMTF por meio do AR de linhas (gráficos (a), (b) e (c)) e do índice *Silhouette* (gráficos (d), (e) e (f)), em experimentos executados sob o conjunto de notícias hiper partidárias com as representações vetoriais binária, TF e TF-IDF



(a) (b) (c)

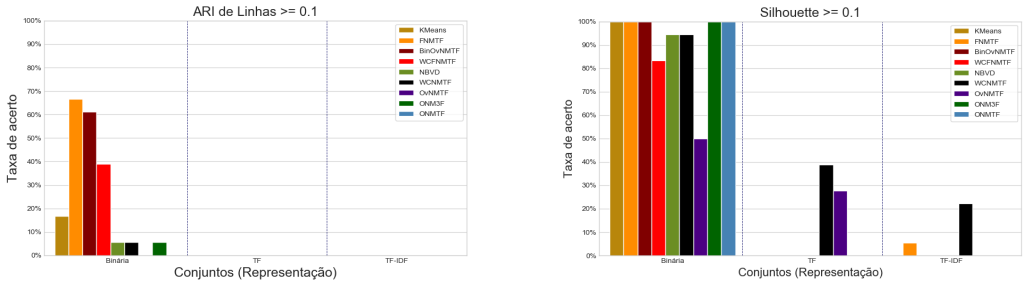
(d) (e) (f)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

A figura 32 apresenta um gráfico de barras que demonstra a taxa de acerto dos algoritmos para o ARI de linhas e para o índice *Silhouette*. O gráfico é semelhante ao que foi discutido na seção 4.1.4, mas avaliando somente os índices acima de 0,1. O gráfico permite sua leitura por meio do exemplo a seguir: na figura 32a o algoritmo FNMTF em pouco mais de 65% das execuções para o conjunto com representação binária atingiu o ARI de linhas maior ou igual a 0,1.

Fica evidente por meio destes gráficos que a representação vetorial binária proporcionou os melhores resultados. As representações TF e TF-IDF não proporcionaram o alcance de nenhum resultado de ARI acima de 0,1. Os algoritmos que possuem restrições binárias na sua especificação (FNMTF, BinOvNMTF, WC-FNMTF e *k-Means*) demonstraram ser melhores que os algoritmos que não possuem tais restrições. Sob a ótica do

Figura 32 – Gráficos comparativos para análise da taxa de acerto dos algoritmos para (a) ARI de linhas e (b) índice *Silhouette* maiores que 0,1



(a) (b)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

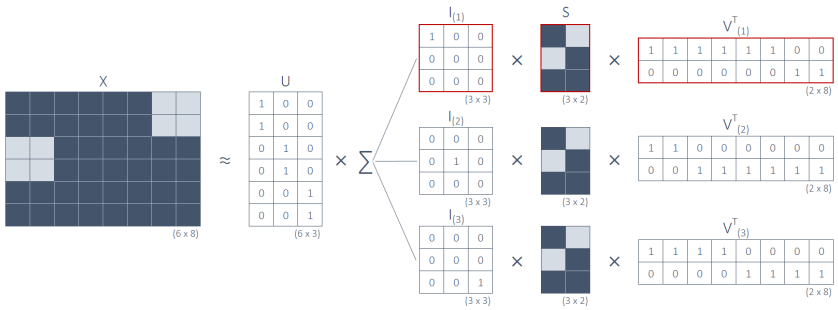
índice *Silhouette*, o gráfico da figura 32b demonstra que cinco dos algoritmos atingiram 100% das execuções para um índice *Silhouette* acima de 0,1. Um fato interessante foi que o algoritmo WC-FNMTF foi o único com restrição binária que atingiu pouco mais de 80% (todos os demais atingiram 100%). Entretanto, ele foi o único algoritmo com restrições binárias que também produziu algum resultado acima de 0,1 para os conjuntos com representações vetoriais TF e TF-IDF. De uma forma geral, o algoritmo WC-FNMTF produziu os resultados mais consistentes para todas as representações deste conjunto de dados.

Por fim, uma das maneiras de avaliar o algoritmo do ponto de vista de análise qualitativa é usando os vetores protótipos gerados pelo processo de fatoração para recuperar as palavras que mais bem representam um grupo de documentos. Os algoritmos com restrições binárias possuem 0 e 1 nas células das matrizes U e V, o que dificulta, no caso de textos, uma avaliação direta dos documentos e palavras que melhor representam tais grupos. Dessa forma, a análise do vetor protótipo pode ser uma alternativa para avaliar os resultados dos algoritmos com restrições binárias. O quadro 2 apresentado na seção 4.1.3 contém a formulação para encontrar os vetores protótipos para cada algoritmo.

As figuras 33 e 34 exemplificam o passo a passo para encontrar os vetores protótipos para o algoritmo BinOvNMTF. Os demais algoritmos seguem a mesma lógica, mas possuem menos matrizes no processo de multiplicação.

A figura 33 (baseada na figura 7) é a ilustração do problema BinOvNMTF por meio de fatoração de matrizes para k igual a 3 e l igual a 2. As marcações em vermelho destacam as matrizes que precisam ser multiplicadas para encontrar o vetor protótipo do grupo 1 de documentos.

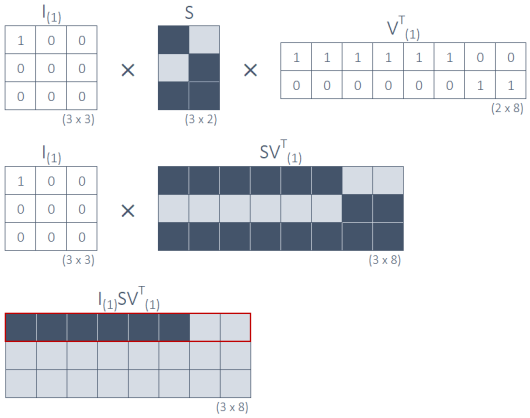
Figura 33 – Ilustração do problema BinOvNMTF para k igual a 3 e l igual a 2, com destaque para as matrizes que são a base do vetor protótipo do grupo 1 de documentos



Fonte: Waldyr Lourenço de Freitas Junior, 2023

A figura 34 ilustra o passo a passo da multiplicação das matrizes $I_{(1)}$, S e $V^T_{(1)}$ para encontrar o vetor protótipo do grupo 1, que é a linha destacada em vermelho (matriz $I_{(1)}SV^T_{(1)}$). As células com a cor azul-escuro caracterizam valores altos na matriz. As palavras que possuem os maiores valores são as que melhor representam o grupo 1.

Figura 34 – Ilustração detalhada da multiplicação das matrizes que são a base do vetor protótipo do grupo 1 de documentos



Fonte: Waldyr Lourenço de Freitas Junior, 2023

O quadro 4 ilustra as 10 principais palavras que melhor representam os grupos de notícias para os algoritmos NBVD, BinOvNMTF, *k-means* e ONM3F. Os dados foram selecionados de uma série de execuções com os parâmetros k igual a 2. A execução de que gerou o maior ARI foi selecionada. A escolha dos algoritmos se baseou no resultado obtido da avaliação humana sobre grupos bem definidos (cf. resultados da atividade 4). A escolha do parâmetro $k = 2$ se baseou na especificidade do corpus, que possui duas classes (notícias hiper partidárias e não hiper partidárias). A interpretação dos resultados é do próprio autor deste trabalho.

Quadro 4 – 10 principais palavras que representam cada grupo gerado pelos algoritmos NBVD, BinOvNMTF, *k-means* e ONM3F

Algoritmo	**Grupo #1**	**Grupo #2**
NBVD	twitter	left
	status	class
	election	ruling
	tweets	ordinary
	type	world
	internal	american
	presidential	article
	september	values
	political	working
	donald	good
Algoritmo	**Grupo #1**	**Grupo #2**
BinOvNMTF	clear	victims
	interview	north
	long	article
	threat	school
	federal	woman
	government	women
	administration	stand
	think	flag
	things	back
	national	children
Algoritmo	**Grupo #1**	**Grupo #2**
K-means	police	left
	donald	money
	twitter	right
	country	police
	american	years
	man	american
	year	world
	fbi	back
	media	think
	women	house
Algoritmo	**Grupo #1**	**Grupo #2**
ONM3F	free	class
	send	ruling
	sms	ordinary
	text	article
	message	good
	islands	fact
	british	california
	republic	found
	kingdon	working
	ireland	view

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Visualmente, é possível perceber que algumas palavras são características de textos hiper partidários (apesar de não definir o grupo como um todo como hiper partidário) e outras não permitem uma definição clara do grupo, como discutido nos tópicos a seguir. As palavras foram analisadas dentro do contexto do grupo e não de maneira isolada, e foi feita uma correlação com a figura 25, explanada na seção 4.2.1.

- **NBVD**: Um grupo não hiper partidário (#1) e o outro que apresenta algumas palavras que podem indicar hiperpartidarismo, a depender do contexto (#2).

- **BinOvNMTF**: Ambos os grupos possuem palavras relacionadas a terrorismo e que, a depender do contexto, podem indicar hiperpartidarismo. Desta forma, ambos os grupos (#1 e #2) estão sem definição.

- ***K-means***: Um grupo com forte tendência a hiperpartidarismo (#1) e outro sem uma definição clara (#2).

- **ONM3F**: Dois grupos sem indicação de hiperpartidarismo (#1 e #2).

As palavras *status* (status), *election* (eleição), *presidential* (presidencial), *political* (política) e *donald* (Donald[8]) sugerem que o grupo #1 do NBVD está relacionado às eleições norte americanas, sem indicação de hiperpartidarismo. De fato, as notícias do corpus datam o período das eleições dos EUA. A palavra *political* aparece na figura 25 como uma das mais frequentes nos textos hiper partidários; a palavra *donald* aparece como palavra frequente tanto nos textos hiper partidários quanto nos que não o são. O outro grupo gerado pelo algoritmo NBVD (#2) trouxe palavras que podem indicar hiperpartidarismo: *left* (esquerda), *class* (classe), *american* (americano) e *values* (valores). As palavras *american* e *left* estão entre as palavras mais frequentes nos textos classificados como hiper partidários.

Ambos os grupos gerados pelo algoritmo BinOvNMTF remetem a terrorismo. Um grupo com menor robustez: *threat* (ameaça), *government* (governo), *national* (nacional), *federal* (federal) - #1 - e outro com maior robustez: *victims* (vítimas), *school* (escola), *women* (mulheres), *children* (crianças) - #2. Neste caso, as notícias podem estar discorrendo sobre terrorismo, mas não necessariamente com tendência hiper partidária pelo autor do texto. Apenas duas palavras (*government* e *women*) foram encontradas entre as mais frequentes, segundo o diagrama da figura 25, ambas nos textos classificados com hiperpartidarismo.

As palavras *police* (polícia), *donald* (Donald), *fbi* (FBI), *man* (homem), *women* (mulheres) do grupo #1 gerado pelo algoritmo *k-means* são as que mais fortemente definem o grupo como hiper partidário, na visão do autor deste trabalho. O contexto é de algum cenário envolvendo o candidato à presidência justamente duramente o período das

[8] Referência ao ex-presidente dos Estados Unidos da América, Donaldo Trump

eleições. Contudo, segundo o diagrama da figura 25, todas as palavras destacam-se entre as mais frequentes dos textos classificados como não hiper partidários, o que corrobora a subjetividade da análise dos grupos. O grupo #2 possui algumas palavras que sugerem hiperpartidarismo, tais como *left* (esquerda), *right*, (direita), *police* (polícia), mas não são o suficiente para dar uma definição clara para o grupo.

De uma forma geral, o cenário demonstra que os algoritmos não puderam separar bem alguns grupos no contexto das notícias, sobretudo para notícias hiper partidárias. Esse comportamento, pelo menos superficialmente, corrobora com os resultados da qualidade do agrupamento dos algoritmos, cujo ARI e *Silhouette* demonstraram não ser tão elevados.

4.2.5 Considerações Finais

De uma forma geral, os algoritmos que possuem restrições binárias apresentaram resultados mais satisfatórios que os que não as possuem, assim como nos experimentos com dados sintéticos. Sob a ótica do índice *Silhouette*, o algoritmo WC-NMTF, proposto no trabalho de Salah, Ailem e Nadif (2018), apresentou os resultados mais robustos para os conjuntos com representação baseada em frequência, no entanto, apresentou também mais dificuldade em reconstruir a matriz original. Os algoritmos que não possuem restrições binárias atingiram os menores erros de reconstrução. Eles possuem números reais nas matrizes U e V, o que favorece a reconstrução.

Além disso, os autores dos algoritmos FNMTF (WANG *et al.*, 2011) e BinOvNMTF (BRUNIALTI *et al.*, 2017) não apresentaram experimentos que explorasse o aspecto de análise qualitativa. Neste trabalho foram realizados experimentos usando os vetores protótipos gerados pelos algoritmos como meio para extrair dados para avaliação humana, uma abordagem nova para algoritmos que possuem restrições binárias. A estratégia de utilizar os vetores protótipos foi observada em alguns estudos da revisão bibliográfica (LEE; SEUNG, 1999; CHEN; WANG; DONG, 2009; ALLAB; LABIOD; NADIF, 2016; AILEM; AGHILES; NADIF, 2017; SALAH; AILEM; NADIF, 2018; CASALINO *et al.*, 2018), mas os algoritmos propostos por eles não possuem restrições binárias. O algoritmo BinOvNMTF foi capaz de capturar palavras que representam notícias hiper partidárias.

O quadro 5 apresenta o resultado consolidado dos experimentos realizados com os dados do mundo real. O “x” representa que o algoritmo teve um bom desempenho nos

experimentos, para as medidas de avaliação relacionadas, ou seja, para o ARI de linhas, índice *Silhouete* e erro de reconstrução, considerando as representações binária, TF e TF-IDF.

Quadro 5 – Quadro consolidado dos experimentos com dados do mundo real

Algoritmo	ARI linhas			*Silhouette*			Erro		
	Bin	TF	TF-IDF	Bin	TF	TF-IDF	Bin	TF	TF-IDF
K-means				x					
NBVD							x	x	x
ONM3F							x	x	x
ONMTF							x	x	x
OvNMTF					x		x	x	x
FNMTF	x			x		x			
BinOvNMTF	x			x					
WC-NMTF					x	x			
WC-FNMTF	x	x	x			x			

Fonte: Waldyr Lourenço de Freitas Junior, 2023

4.3 Avaliação qualitativa de textos com alunos de graduação

Esta seção apresenta os experimentos realizados com participação de humanos na análise qualitativa dos resultados produzidos pelos algoritmos sob avaliação. O corpus utilizado foi o mesmo apresentado na seção 4.2.1. A seção 4.3.2 discute cada uma das atividades que foram aplicadas com interação humana. A seção 4.3.3 apresenta uma discussão dos resultados decorrentes dos experimentos. Por fim, a seção 4.3.4 discute as considerações finais.

Tais experimentos[9] foram realizados com alunos de duas disciplinas do curso de sistemas de informação da Escola de Artes, Ciências e Humanidades (EACH) da Universidade de São Paulo (USP). As disciplinas foram ACH2016 - Inteligência Artificial e ACH2187 - Mineração de Dados. Ao todo, 118 alunos diferentes participaram da pesquisa, sendo 61 da disciplina de Inteligência Artificial e 57 da disciplina de Mineração de Dados. Eles participaram de quatro atividades ao longo da disciplina, conforme detalhado na seção 4.3.2.

[9] Para a realização de experimentos com a participação de pessoas, foi apresentado um projeto ao Comitê de Ética em Pesquisa (CEP) da Universidade de São Paulo (USP) detalhando as atividades que seriam realizadas. O projeto foi aprovado pelo CEP no dia 14 de julho de 2021.

4.3.1 Conjunto de dados

O corpus utilizado para as atividades realizadas com os alunos foi o mesmo já detalhado na seção 4.2.1.

4.3.2 Atividades

Ao todo, foram realizadas quatro atividades distintas ao longo da ministração das disciplinas. Em cada uma das atividades, foi apresentada aos alunos a definição de notícia hiper partidária. Essa definição foi estabelecida pelos próprios autores do corpus, a saber: *Uma notícia que segue uma argumentação hiper partidária exibe lealdade cega, preconceituosa ou irracional a uma parte, facção, causa ou pessoa.* Cada uma das atividades foi detalhada nos tópicos a seguir.

Atividade 1

A atividade 1 foi formada por duas tarefas. Na primeira, os alunos precisavam ler uma notícia do corpus e classificá-la entre hiper partidária e não hiper partidária. Em seguida, eles deviam selecionar as 10 palavras mais relevantes que justificassem a escolha deles e então ordená-las, da mais relevante para a menos relevante. Na segunda tarefa, uma notícia já rotulada (hiper partidária ou não hiper partidária) foi apresentada aos alunos. Eles também tinham que selecionar as 10 palavras mais relevantes para representar o rótulo da notícia e ordená-las do mesmo modo da tarefa 1.

Para esta atividade, o corpus foi pré-processado com as representações vetoriais binária, TF e TF-IDF. Para cada representação vetorial, as duas notícias mais próximas dos centroides dos grupos foram selecionadas para as tarefas, uma hiper partidária e outra não hiper partidária, de acordo com a rotulação original do corpus. Além disso, ainda para cada representação, enquanto um grupo de alunos iniciou a tarefa 1 com uma notícia hiper partidária, o outro iniciou a mesma tarefa com uma notícia não hiper partidária. Dessa forma, a mesma atividade foi realizada dividindo o grupo de alunos em seis, com o objetivo de mitigar qualquer viés.

O objetivo desta atividade foi realizar uma análise das palavras que mais bem representavam os grupos, de acordo com a visão dos alunos, a fim de permitir uma comparação com os resultados dos algoritmos.

Atividades 2 e 3

As atividades 2 e 3 foram formadas por três tarefas cada. No início de cada tarefa, para cada uma das atividades, um quadro contendo diversos grupos de palavras em língua inglesa era apresentado aos alunos (um único quadro foi apresentado nas três tarefas da atividade 2 e um outro quadro foi apresentado nas três tarefas da atividade 3). Cada grupo nestes quadros continha 10 palavras, que foram extraídas do processamento de cada um dos grupos de cada um dos algoritmos. Para a atividade 2, o quadro continha 25 grupos de palavras, conforme figura 35. Para montar este quadro, foram selecionados os parâmetros dos melhores resultados de cada algoritmo para o ARI e Índice *Silhouette*. Em seguida, foram feitas 10 novas execuções dos algoritmos com esses parâmetros e escolhida a melhor execução de cada um deles. Por meio das matrizes fatoradas, foram selecionadas as palavras que mais bem representavam os grupos gerados pelos algoritmos, de acordo com a heurística de uso dos vetores protótipos. Para a atividade 3, o quadro continha 30 grupos de palavras, conforme apresentado na figura 36. Para esta atividade foram selecionados os parâmetros dos piores resultados dos índices de validação. Também foram feitas 10 novas execuções dos algoritmos com esses parâmetros e escolhida a pior execução de cada um deles. A ideia foi contrastar o resultado quantitativo dos algoritmos com a percepção dos alunos na avaliação dos grupos. Além disso, intencionalmente, os alunos não tinham conhecimento de quais algoritmos geraram quais grupos.

Especificamente para a atividade 3[10], cinco grupos artificiais foram colocados para efeito de controle da qualidade das respostas, por esta razão há um total de 30 grupos. Os grupos artificiais estão em destaque na figura 36, e foram criados pelo autor deste trabalho para representar fortemente *política* (grupo [5,3]), *esporte* (grupo [2,1]), *stopwords* (grupos [3,4] e [4,2]) e *hiperpartidarismo* (grupo [6,1]).

Na primeira tarefa, os alunos precisavam rotular cada grupo de palavras com base em sete rótulos: Cultura, Educação, Esporte, Negócios, Política, Saúde e Tecnologia. Na

[10] Somente após as análises preliminares da atividade 2 é que foi identificada a necessidade de incluir os grupos controlados nas atividades seguintes com os alunos

segunda tarefa, os alunos precisavam rotular os mesmos grupos com base em dois rótulos: hiper partidário e não hiper partidário. Por fim, na última tarefa eles precisavam responder se a informação contida em cada grupo de palavras foi suficiente ou insuficiente para rotulá-lo como hiper partidário ou não hiper partidário.

Figura 35 – Quadro com os 25 grupos de palavras utilizado nas tarefas da atividade 2

	1		2		3		4		5	
1	joe	elect	trusted	amounts	rule	offices	car	government	clare	university
	principles	living	plays	isis	holiday	islamist	republican	part	roger	god
	google	gonna	star	mean	meant	threats	republicans	especially	oppose	eyes
	goods	live	hours	america	carter	countless	party	espn	arrives	van
	elected	undocumented	divorce	francisco	grateful	particular	create	supremacist	winning	calls
2	wormer	exact	turned	slightly	trusted	divorce	podcast	moms	hole	hitting
	federal	behalf	howard	terrorist	hours	situation	american	fought	hour	sale
	killer	wonder	statement	mesquite	star	term	wrongdoing	rico	poor	authorities
	points	goldwater	freddie	elected	france	mental	america	dominant	back	safety
	became	grateful	selling	tumor	seeking	efforts	male	measures	historic	hollywood
3	clark	wish	rule	offices	car	government	stand	truck	writer	call
	room	skip	holiday	islamist	republican	part	hotel	praying	tillerson	sticking
	opposition	vehicles	meant	threats	republicans	especially	effort	irresponsible	reopened	david
	updated	district	carter	countless	party	espn	troops	canada	offices	forms
	calm	roll	grateful	particular	create	supremacist	intended	players	rent	baby
4	led	american	clark	art	led	camp	offer	email	father	efforts
	class	article	room	gold	class	wondered	commented	elizabeth	wind	camera
	rubin	usa	america	face	rubin	small	lombardo	unified	wounded	pleased
	options	wondered	opposition	wish	options	vetting	undocumented	moore	america	terry
	woods	facing	witnessed	updated	usa	route	presence	living	main	witnessed
5	tuesday	calif	divisive	troops	plays	america	rule	offices	frame	irony
	round	owner	selfish	person	trusted	mean	holiday	islamist	section	repeat
	leaving	terrifying	channel	libya	france	naked	meant	threats	sister	kill
	starting	housing	changing	global	ticket	terry	carter	countless	teneo	bring
	frankly	andrew	trigger	able	main	cabinet	grateful	particular	meme	involved

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Figura 36 – Quadro com os 30 grupos de palavras utilizado nas tarefas da atividade 3

	1		2		3		4		5	
1	bottle	daily	become	constitution	marriage	hickory	exonerating	causal	harvest	compensate
	kurtz	dignity	alhaji	lapse	moretz	manhunt	catastrophic	gearing	festival	handedness
	hidalgo	noting	hold	djungarian	nato	eisenhower	lines	cause	bradpaisley	movie
	animal	begala	kneeling	irrelevance	birth	bloomberg	generation	causing	industry	denuclearisation
	orlando	lose	otherwise	departure	appropriate	hide	caught	garcetti	draftkings	marxism
2	stadium	exercise	filing	noord	depicting	evening	gingrich	genuinely	harvest	hatreon
	goal	rule	bought	animals	legalization	barrett	melania	batten	festival	painted
	plays	america	lack	outdated	looks	karate	bodies	direction	movie	draftkings
	training	gold	highways	nuclearised	molehill	fallout	hospitality	led	marxism	freebiealert
	woods	athlete	outlets	diners	genital	evasion	daughters	nevada	option	concluded
3	harvest	compensate	marriage	hickory	exonerating	causal	anywhere	regardless	hindu	mice
	festival	master	moretz	manhunt	catastrophic	gearing	beside	should	deleted	democratic
	film	boundaries	nato	eisenhower	lines	cause	particular	able	article	herb
	hunt	bedlam	birth	bloomberg	generation	causing	however	small	hacking	leaders
	neighborhoods	aliens	appropriate	hide	caught	garcetti	term	instead	countless	betrayal
4	necessarily	individual	behalf	small	hate	gingrich	gooding	craigslist	becoming	latina
	dynamic	equivlancy	will	such	bloomberg	missing	legislative	mysticism	homeland	islam
	marvel	necessary	himself	thereby	explanation	mohammed	generation	mishandled	komedi	depicting
	compassionate	letting	bring	very	jewish	delaney	marine	caught	outrageous	focused
	controversy	divorce	another	particular	ground	anarchism	megaphone	batman	construction	drive
5	harvest	compensate	harvest	draftkings	party	federal	gooding	craigslist	harvest	master
	festival	handedness	festival	impetus	law	ideology	legislative	mysticism	festival	aliens
	hunt	movie	hunt	option	changing	principles	generation	mishandled	film	lane
	industry	denuclearisation	industry	hatreon	government	election	marine	caught	boundaries	compensate
	draftkings	marxism	bradpaisley	handedness	public	democracy	megaphone	mean	hunt	hook
6	e-mail	unified	gingrich	genuinely	morning	marine	deportation	hamburg	mark	indian
	victims	white	melania	batten	batman	defuse	lemon	emotion	dutchman	events
	opposition	person	bodies	direction	gearing	caught	money	machine	control	immanuel
	trusted	anti	hospitality	led	assassinations	gooding	european	negotiations	natural	half
	censure	sexism	daughters	nevada	countless	idea	family	head	equal	behind

Fonte: Waldyr Lourenço de Freitas Junior, 2023

O objetivo destas atividades foi avaliar, sob uma ótica de interpretação humana, a capacidade dos algoritmos em agrupar dados e avaliar a qualidade da informação produzida por eles. A atividade 3 também objetivou encontrar uma relação entre os resultados mal

avaliados segundo medidas quantitativas e a interpretação dos grupos de palavras feita pelos alunos.

Atividade 4

A quarta e última atividade com os alunos foi formada por uma única tarefa. Nesta tarefa, um novo quadro com diversos grupos de palavras foi apresentado aos alunos. Cada grupo também continha 10 palavras, que mais bem representavam os grupos gerados pelos algoritmos. Havia uma referência aos algoritmos (Algoritmo 1, Algoritmo 2, ..., Algoritmo 10), sem nomeá-los. Para cada algoritmo, foram selecionados os parâmetros que proporcionaram o maior valor de ARI e então foram realizadas 10 novas execuções, para cada algoritmo, com estes parâmetros. A melhor execução foi escolhida, e, de acordo com a heurística de uso da matriz S e dos vetores protótipos gerados durante o processo de fatoração, foram selecionadas as palavras mais representativas para cada grupo. Os alunos precisavam apontar quais eram os grupos mais bem definidos e distintos, e ordenar os algoritmos do melhor para o pior. Alguns grupos de palavras controlados também foram inseridos no quadro, associado a um algoritmo *fake*, para avaliar a eficácia das respostas dos alunos. Um exemplo do quadro utilizado está demonstrado na figura 37, com destaque para os grupos controlados[11]. Essa atividade foi dividida em três grupos de alunos, para explorar resultados com k igual a 2, 3 e 4.

O objetivo desta atividade foi avaliar os algoritmos do ponto de vista qualitativo, baseado na percepção dos alunos em grupos mais bem definidos e distintos.

4.3.3 Resultados

O gráfico na figura 38 apresenta o resultado obtido a partir da realização da tarefa 1 para a atividade 1. No eixo x estão divididos os seis grupos de alunos. Cada setor do gráfico apresenta a classificação feita pelos alunos, em notícias rotuladas como hiper partidárias ("Hiper") e não hiper partidárias ("Não"). Os rótulos verdadeiros foram destacados em verde-claro, para facilitar a análise. O eixo y apresenta o percentual da classificação.

A maior parte dos alunos dos grupos 1 e 3 classificaram as notícias com o rótulo não esperado. Quase todos os alunos do grupo 2 escolheram o rótulo verdadeiro; foi o único caso

[11] Os demais quadros utilizados na atividade 4 estão no apêndice B

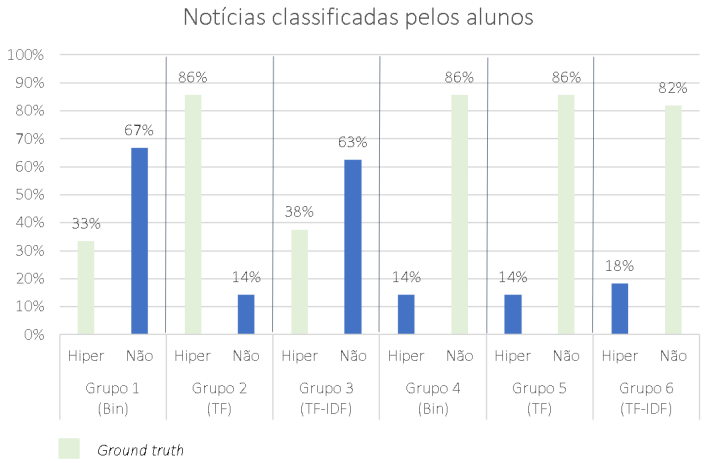
Figura 37 – Quadro com os 10 grupos de palavras utilizado na atividade 4 para $k = 2$

	Grupo 1		Grupo 2	
Algoritmo 1	police	man	left	american
	donald	year	money	world
	twitter	fbi	right	back
	country	media	police	think
	american	women	years	house
Algoritmo 2	cause	future	cause	future
	spoke	speaking	spoke	speaking
	game	monday	game	monday
	speech	money	speech	money
	special	million	special	million
Algoritmo 3	clear	government	victims	women
	interview	administration	north	stand
	long	think	article	flag
	threat	things	school	back
	federal	national	woman	children
Algoritmo 4	data	century	data	century
	found	florida	found	florida
	site	simply	site	simply
	center	mexico	center	mexico
	single	similar	single	similar
Algoritmo 5	twitter	internal	left	american
	status	presidential	class	article
	election	september	ruling	values
	tweets	political	ordinary	working
	type	donald	world	good

	Grupo 1		Grupo 2	
Algoritmo 6	twitter	year	trump	people
	people	election	president	white
	obama	state	clinton	obama
	police	back	hillary	campaign
	news	october	donald	america
Algoritmo 7	free	islands	class	fact
	send	british	ruling	california
	sms	republic	ordinary	found
	text	kingdom	article	working
	message	ireland	good	view
Algoritmo 8	class	american	free	islands
	twitter	world	send	ireland
	status	ordinary	sms	republic
	ruling	article	text	indian
	message	good	message	netherlands
Algoritmo 9	way	believe	way	political
	called	political	called	believe
	never	public	long	work
	long	day	never	public
	fact	work	fact	good
Algoritmo 10	e-mail	unified	party	federal
	victims	white	law	ideology
	opposition	person	changing	principles
	trusted	anti	government	election
	censure	sexism	public	democracy

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Figura 38 – Resultado da classificação das notícias feita pelos alunos como parte da tarefa 1 da atividade 1



Fonte: Waldyr Lourenço de Freitas Junior, 2023

em que isso aconteceu para o rótulo hiper partidário. As principais palavras selecionadas por eles para representar essa notícia hiper partidária foram *brainwashing* (lavagem cerebral), *victimization* (vitimização), *dependency* (dependência), *rag-tag* (gentalha), *conservative* (conservador). Todas essas palavras possuem uma frequência baixa no corpus e elas não foram selecionadas pelos algoritmos, nas demais atividades, como mais representativas dos grupos de notícias.

Os alunos dos grupos 4, 5 e 6 classificaram as notícias conforme esperado; as notícias de fato não eram hiper partidárias (segundo a classificação do autor do corpus). Contudo,

observa-se que houve uma tendência em classificar as notícias como não hiper partidárias. O uso do idioma inglês pode ter dificultado a realização da tarefa, embora tenha sido recomendado que os alunos utilizassem aplicativos de tradução de idioma, em caso de dúvidas[12]. Uma outra causa é que as notícias são predominantemente de política e a temática de hiperpartidarismo normalmente está associada à política. Isso pode tê-los deixado com dúvidas se de fato a notícia era hiper partidária ou somente uma notícia sobre política. Além disso, os alunos dos grupos 2 e 5, que avaliaram as notícias que foram selecionadas por meio de um conjunto pré-processado com a representação vetorial TF, acertaram o rótulo verdadeiro em 86%.

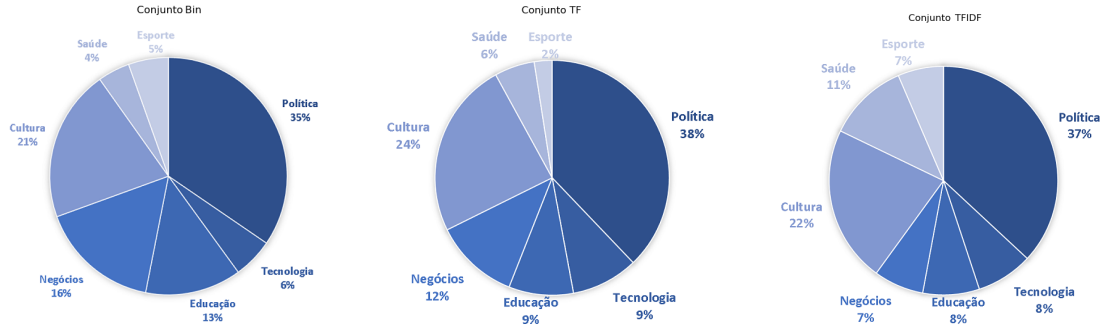
Para a tarefa 2 da atividade 1, como já explanado anteriormente, as notícias já foram apresentadas rotuladas. As principais palavras escolhidas pelos participantes para representar as notícias hiper partidárias foram *feared* (temida(o)), *apology* (desculpa), *accused* (acusada(o)), *raped* (estuprada(o)), *irresponsible* (irresponsável), *pathetic* (patética(o)), *disgust* (desgosto), *corrupt* (corrupta(o)). Essas palavras encontram-se entre as com frequências baixa e média no corpus, e não aparecem, por exemplo, no diagrama apresentado na figura 25. A palavra *irresponsible* aparece mais nos textos rotulados com hiperpartidarismo, corroborando com a avaliação dos alunos. Essa palavra também foi uma das identificadas pelos algoritmos como mais representativa dos grupos gerados por eles (grupos gerados para Atividade 2). De uma forma geral, a atividade 1 demonstrou pouca relação entre as palavras escolhidas pelos alunos e as palavras mais frequentes do corpus ou as que mais representavam os grupos gerados pelos algoritmos.

Os gráficos apresentados na figura 39 dão uma visão geral dos resultados da tarefa 1 da atividade 2, em que era pedido que os alunos classificassem os grupos de palavras extraídos dos resultados dos algoritmos. Mais da metade dos alunos rotularam os grupos de palavras como pertencentes à política e cultura. Em seguida, os rótulos mais escolhidos foram negócios e educação.

Não havia conhecimento prévio, por parte dos alunos, sobre quais algoritmos geraram quais grupos de palavras. A informação foi utilizada a posteriori para avaliar o desempenho de cada algoritmo. Por exemplo, suponha que 90% dos alunos classificaram um grupo de palavras como sendo de política, isso indica que o algoritmo que gerou este grupo teve êxito em separar as palavras que representavam política.

[12] Não foi realizada nenhuma avaliação de proficiência do idioma com os alunos

Figura 39 – Gráficos que apresentam a classificação relativa dos grupos de palavras gerados pelos algoritmos para a tarefa 1 da atividade 2, para os conjuntos com representação (a) binária, (b) TF e (c) TF-IDF.



(a) (b) (c)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

A tabela 5 apresenta essa análise. Os números representam a quantidade de vezes que determinado algoritmo gerou um grupo em que mais de 50% dos alunos o classificaram para o mesmo rótulo. Como exemplo, o algoritmo BinOvNMTF, representação TF com índice *Silhouette*, foi responsável por gerar três grupos em que mais de 50% dos alunos o classificaram para o mesmo rótulo, de um total de 4 possíveis.

Tabela 5 – Contagem do número de vezes em que os algoritmos geraram grupos em que mais de 50% dos respondentes da tarefa 1 da atividade 2 os classificaram para um mesmo rótulo

Algoritmo	Silhouette			ARI	TOTAL
	Bin	TF	TF-IDF		
K-means	1/4	1/3	0/0	0/4	2/11
NBVD	0/2	1/2	2/4	1/2	4/10
ONM3F	2/2	1/2	2/4	1/2	6/10
ONMTF	2/4	2/3	1/3	0/0	5/10
OvNMTF	1/2	3/3	2/4	2/4	8/13
FNMTF	1/4	0/3	3/3	0/4	4/14
BinOvNMTF	2/3	3/4	1/3	2/3	8/13
WC-NMTF	0/2	1/2	1/2	0/4	2/10
WC-FNMTF	2/2	3/3	0/2	0/2	5/9

Fonte: Waldyr Lourenço de Freitas Junior, 2023

De uma forma geral, os algoritmos BinOvNMTF e OvNMTF tiveram o melhor resultado, seguidos por ONM3F, WC-FNMTF e ONMTF. Essa análise permite indicar que tais algoritmos, sob uma ótica humana, conseguem gerar grupos cujas palavras estão mais bem relacionadas ao rótulo desse grupo. Para aumentar a acurácia da análise, a mesma avaliação foi feita aumentando o percentual para 70% e pode ser visualizada na tabela 6. Neste caso, o algoritmo WC-FNMTF se destacou com o melhor resultado.

Tabela 6 – Contagem do número de vezes em que os algoritmos geraram grupos em que mais de 70% dos respondentes da tarefa 1 da atividade 2 os classificaram para um mesmo rótulo

Algoritmo	Silhouette			ARI	TOTAL
	Bin	TF	TF-IDF		
K-means	1/4	0/3	0/0	0/4	1/11
NBVD	0/2	1/2	0/4	0/2	1/10
ONM3F	0/2	0/2	1/4	1/2	2/10
ONMTF	1/4	0/3	0/3	0/0	1/10
OvNMTF	0/2	0/3	1/4	0/4	1/13
FNMTF	0/4	0/3	1/3	0/4	1/14
BinOvNMTF	1/3	1/4	0/3	0/3	2/13
WC-NMTF	0/2	0/2	0/2	0/4	0/10
WC-FNMTF	2/2	3/3	0/2	0/2	5/9

Fonte: Waldyr Lourenço de Freitas Junior, 2023

A mesma avaliação foi realizada para a tarefa 2 da atividade 2, em que os alunos precisavam classificar os grupos de palavras com base nos rótulos do próprio corpus, hiper partidário e não hiper partidário. Neste caso, como foram apenas dois rótulos, o percentual escolhido foi de 80% e 90%. A tabela 7 apresenta os resultados fixando o percentual de 80% e a tabela 8 faz o mesmo para 90%.

O algoritmo WC-FNMTF se destacou para o percentual de 80%, em seguida os algoritmos ONM3F, OvNMTF, ONMTF e WC-NMTF apresentaram os melhores resultados. Para o percentual de 90%, o algoritmo WC-FNMTF teve ligeiramente a maior contagem e melhor posição relativa, seguida dos algotimos ONMTF e *k-means*. Observou-se também nesta tarefa uma tendência dos respondentes em classificar os grupos como não hiper partidários.

De maneira geral, os algoritmos com restrições binárias que se destacaram foram o WC-FNMTF e BinOvNMTF. Dentre os que não possuem restrições, os algoritmos ONM3F, ONMTF e OvNMTF se destacaram. De uma forma geral, foi possível avaliar a capacidade dos algoritmos em agrupar dados e a qualidade dos grupos gerados por eles por meio da análise dos alunos.

Os gráficos apresentados na figura 40 são da tarefa 1 da atividade 3. Nesta tarefa foi pedido que os alunos classificassem alguns grupos gerados pelos algoritmos, e assim como foi observado na tarefa da atividade 2, os alunos rotularam os grupos de palavras como pertencentes à política e cultura, mas houve uma tendência maior em classificar os

Tabela 7 – Contagem do número de vezes em que os algoritmos geraram grupos que mais de 80% dos respondentes da tarefa 2 da atividade 2 os classificaram para um mesmo rótulo

Algoritmo	Silhouette			ARI	TOTAL
	Bin	TF	TF-IDF		
K-means	2/4	0/3	0/0	1/4	3/11
NBVD	1/2	0/2	1/4	0/2	2/10
ONM3F	1/2	0/2	3/4	0/2	4/10
ONMTF	0/4	1/2	2/3	0/0	3/10
OvNMTF	2/2	0/3	0/4	2/4	4/13
FNMTF	1/4	0/3	0/3	1/4	2/14
BinOvNMTF	2/3	1/4	0/3	0/3	3/13
WC-NMTF	2/2	0/2	1/2	0/4	3/10
WC-FNMTF	2/2	3/3	1/2	1/2	7/9

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Tabela 8 – Contagem do número de vezes em que os algoritmos geraram grupos que mais de 90% dos respondentes da tarefa 2 da atividade 2 os classificaram para um mesmo rótulo

Algoritmo	Silhouette			ARI	TOTAL
	Bin	TF	TF-IDF		
K-means	2/4	0/3	0/0	0/4	2/11
NBVD	0/2	0/2	0/4	0/2	0/10
ONM3F	1/2	0/2	0/4	0/2	1/10
ONMTF	0/4	1/3	1/3	0/0	2/10
OvNMTF	0/2	0/3	0/4	0/4	0/13
FNMTF	0/4	0/3	0/3	0/4	0/14
BinOvNMTF	0/3	1/4	0/3	0/3	1/13
WC-NMTF	1/2	0/2	0/2	0/4	1/10
WC-FNMTF	0/2	2/3	1/2	0/2	3/9

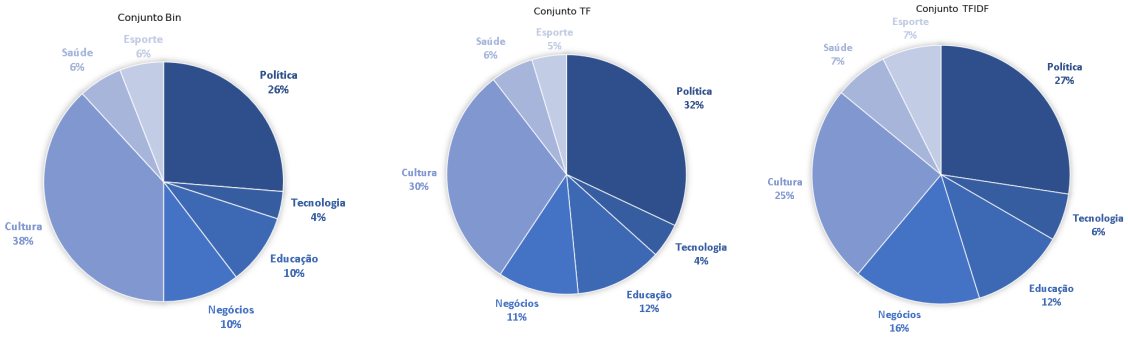
Fonte: Waldyr Lourenço de Freitas Junior, 2023

grupos como cultura. Após estes, os rótulos negócios e educação foram os mais escolhidos, em uma proporção relativamente maior que a atividade 2.

As notícias do corpus são predominantemente sobre política e a tendência dos alunos em escolher outros rótulos sugerem que as palavras dos grupos gerados pelos algoritmos nesta tarefa não foram suficientes para indicar um rótulo predominante. Importante destacar que esta tarefa utilizou as piores execuções dos algoritmos e os grupos gerados por eles não foram assertivos. A grande maioria dos alunos rotularam os grupos controlados para o rótulo verdadeiro, o que aumenta a confiança nas respostas dos alunos.

A tabela 9 apresenta a análise de contagem para a tarefa 2 da atividade 3. Ela apresenta a visão em que mais de 80% dos alunos classificaram os grupos para o mesmo rótulo (hiper partidário e não hiper partidário). Os resultados demonstraram que o

Figura 40 – Gráficos que apresentam a classificação relativa dos grupos de palavras gerados pelos algoritmos para a tarefa 1 da atividade 3, para os conjuntos com representações (a) binária, (b) TF e (c) TF-IDF.



(a) (b) (c)

Fonte: Waldyr Lourenço de Freitas Junior, 2023

algoritmo *k-means* obteve o melhor resultado, seguido do ONMTF. Estes algoritmos não apresentaram resultados representativos para a tarefa semelhante da atividade 2, comportamento que sugere certa arbitrariedade. Aparentemente, o resultado quantitativo dos algoritmos influencia a avaliação qualitativa realizada sobre seus resultados.

Tabela 9 – Contagem do número de vezes em que os algoritmos geraram grupos que mais de 80% dos respondentes da tarefa 2 da atividade 3 os classificaram para um mesmo rótulo

Algoritmo	Silhouette			ARI	TOTAL
	Bin	TF	TF-IDF		
K-means	1/2	3/4	1/3	0/2	5/11
NBVD	1/4	1/4	1/2	0/4	3/14
ONM3F	1/3	1/3	0/4	1/3	3/13
ONMTF	0/2	2/2	1/3	1/3	4/10
OvNMTF	0/4	2/2	1/2	0/3	3/11
FNMTF	0/2	0/2	1/2	0/2	1/8
BinOvNMTF	1/2	1/2	0/4	1/3	3/11
WC-NMTF	0/4	0/4	0/2	0/3	0/13
WC-FNMTF	0/2	0/2	2/3	0/2	2/9

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Os resultados da atividade 4 são demonstrados por meio de um mapa de calor em tons de amarelo e verde. Uma contagem alta terá a cor verde mais escura e uma contagem mais baixa terá a cor amarela mais clara. Na visão da tarefa, a contagem representa o número de vezes que os alunos escolheram determinado grupo de palavras como sendo o mais bem definido e distinto dos demais. Na prática, isso representa o número de vezes que um algoritmo foi escolhido em determinada posição.

A figura 41 demonstra os resultados dos experimentos para $k = 2$. Ao avaliar o *rank* dos três melhores, observa-se que os algoritmos NBVD, BinOvNMTF, ONM3F e o *Fake* (grupo controlado) tiveram os melhores resultados. O algoritmo NBVD foi o mais bem ranqueado, o que significa que os respondentes elegeram seus grupos como os mais bem definidos e distintos. Era esperado que o algoritmo *Fake*, que traz alguns grupos gerados manualmente para efeito de controle, fosse ranqueado como o melhor. Possivelmente, o número de grupos baixo não permitiu aos alunos distingui-lo facilmente dentre os demais. Outra observação interessante é que os algoritmos FNMTF, WC-NMTF e WC-FNMTF apresentaram os piores resultados.

Figura 41 – Mapa de calor para representar o resultado do ranqueamento dos algoritmos para a atividade 4 com $k = 2$

K = 2

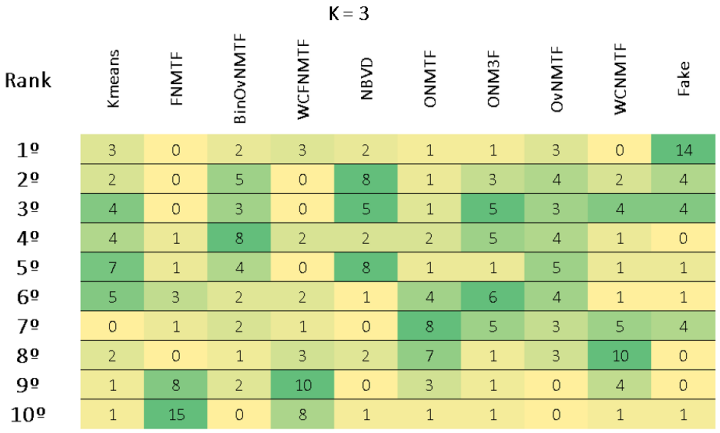
Rank	Kmeans	FNMTF	BinOvNMTF	WCFNMTF	NBVD	ONMTF	ONM3F	OvNMTF	WCNMTF	Fake
1º	2	1	4	0	8	3	3	1	0	3
2º	2	0	4	0	6	3	5	1	0	4
3º	5	0	3	1	4	2	3	3	0	4
4º	6	0	6	1	2	0	4	1	1	4
5º	2	0	5	0	2	2	2	5	2	5
6º	3	1	0	1	1	8	1	5	1	4
7º	4	1	2	1	0	7	1	6	2	1
8º	0	6	0	1	1	0	5	1	11	0
9º	0	11	0	10	1	0	1	1	1	0
10º	1	5	1	10	0	0	0	1	7	0

Fonte: Waldyr Lourenço de Freitas Junior, 2023

A figura 42 apresenta a mesma visão, mas para $k = 3$. Dessa vez o algoritmo *Fake* foi ranqueado como o melhor, o que realmente era esperado. Mais uma vez os algoritmos NBVD e BinOvNMTF tiveram os melhores resultados. Isso pode indicar que os algoritmos têm uma capacidade de gerar grupos mais bem definidos e distintos segundo a percepção humana. O mesmo comportamento para os algoritmos FNMTF, WC-NMTF e WC-FNMTF foi observado neste experimento. O algoritmo FNMTF teve uma queda considerável no desempenho e ficou como o pior ranqueado.

A figura 43 apresenta os resultados para $k = 4$. O algoritmo *Fake* novamente foi o melhor ranqueado, o que indica que à medida que o k aumenta, ou seja, mais grupos são apresentados para as pessoas, ficou mais fácil de distingui-los; essa observação também aumenta a confiança nas respostas dos alunos. O algoritmo BinOvNMTF apresentou uma queda de desempenho para um valor maior de k. Outra observação interessante foi o comportamento do WC-FNMTF, que melhorou consideravelmente seu desempenho e ficou

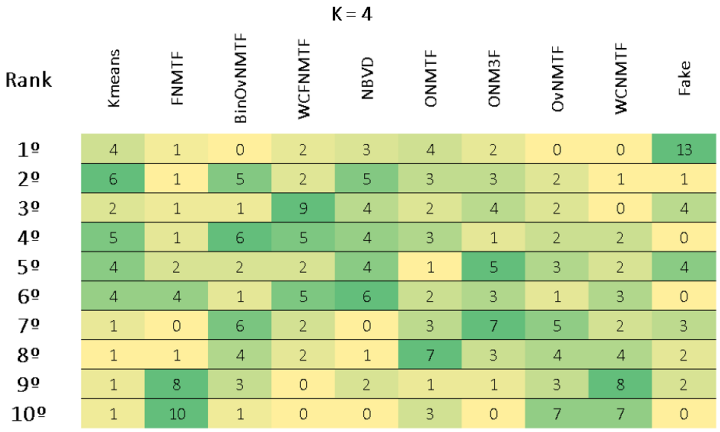
Figura 42 – Mapa de calor para representar o resultado do ranqueamento dos algoritmos para a atividade 4 com $k = 3$



Fonte: Waldyr Lourenço de Freitas Junior, 2023

ranqueado entre os melhores. De uma forma geral, o algoritmo NBVD ainda permanece ranqueado como o melhor. Os quadros com as palavras da atividade 4 para $k = 3$ e $k = 4$ podem ser visualizadas no apêndice B.

Figura 43 – Mapa de calor para representar o resultado do ranqueamento dos algoritmos para a atividade 4 com $k = 4$



Fonte: Waldyr Lourenço de Freitas Junior, 2023

4.3.4 Considerações Finais

Sob a ótica de avaliação humana, o algoritmo WC-FNMTF apresentou os melhores resultados para as tarefas que envolviam classificação de grupos. Suas principais características são o uso de uma matriz de coocorrência de palavras, para preservação da semântica que há entre elas, e as restrições binárias. O algoritmo ONM3F também apre-

sentou bons resultados nas tarefas de classificação dos grupos e sua principal característica é a restrição de ortogonalidade.

De uma forma geral, não foi identificada uma relação forte entre as palavras escolhidas pelos alunos para rotular as notícias e as palavras mais frequentes para este rótulo no corpus. Sobre as tarefas que os alunos precisavam indicar se a informação dos grupos de palavras era suficiente ou não para rotular tais grupos como hiper partidários ou não, observou-se uma certa divisão entre eles. De uma forma geral, essa observação não permitiu inferir algo nessa tarefa.

Por fim, o algoritmo NBVD, que é o algoritmo tradicional de fatoração tripla de matrizes não negativas, foi superior na análise de grupos distintos e bem definidos. Os algoritmos com restrições binárias tiveram melhores resultados para k maior, com exceção do FNMTF que foi ranqueado dentre os piores para todos os valores de k.

O quadro 6 apresenta o resultado consolidado da análise qualitativa realizada com os alunos. O "x" representa que o algoritmo teve um bom desempenho para o experimento relacionado.

Quadro 6 – Quadro consolidado da análise qualitativa realizada com pessoas

Algoritmo	Análise quantitativa com pessoas	
	Classificação	Grupos definidos e distintos
K-means	x	
NBVD		x
ONM3F	x	
ONMTF		
OvNMTF		
FNMTF		
BinOvNMTF	x	x
WC-NMTF		
WC-FNMTF	x	

Fonte: Waldyr Lourenço de Freitas Junior, 2023

4.4 Considerações finais gerais

O quadro 7 apresenta uma visão consolidada dos experimentos realizados, sob a ótica de análise quantitativa e qualitativa. O "x" representa que o algoritmo teve um bom desempenho para o experimento relacionado. A visão é baseada nos índices de validação ARI e *Silhouette* (SIL), erro de reconstrução da matriz original (Erro), tempo de execução (Tempo) e análise das pessoas (Pessoas).

Os algoritmos WC-FNMTF, FNMTF e BinOvNMTF se destacaram na maior parte dos experimentos. O BinOvNMTF e o FNMTF demonstraram baixa capacidade para reconstruir a matriz original, tanto para dados sintéticos quanto para dados do mundo real. Os algoritmos sem restrições binárias reconstroem bem a matriz original, ou seja, atingem baixo erro de reconstrução. Nos experimentos com os alunos, WC-FNMTF e BinOvNMTF novamente se destacaram, seguidos dos algoritmos NBVD e ONM3F, que não possuem restrições binárias.

Quadro 7 – Quadro consolidado de todos os experimentos

Algoritmo	Análise quantitativa							Análise qualitativa
	Dados sintéticos			Dados do mundo real				Dados do mundo real
	ARI	Erro	Tempo	ARI	SIL	Erro	Tempo	Pessoas
K-means								x
NBVD						x	x	x
ONM3F	**x		x			x		x
ONMTF						x		
OvNMTF						x		
FNMTF	x		x	x	x		x	
BinOvNMTF	x			x	x			x
WC-NMTF	x				x			
WC-FNMTF	x	x	x	x	x		x	x

** Para conjuntos com esparsidade alta

Fonte: Waldyr Lourenço de Freitas Junior, 2023

As tabelas 10 e 11 apresentam uma visão consolidada do tempo médio de execução dos algoritmos de coagrupamento[13]. A tabela 10 apresenta os resultados para os conjuntos de dados sintéticos e a tabela 11 apresenta os resultados para os conjuntos com dados do mundo real.

Os algoritmos WC-FNMTF e FNMTF possuem o menor tempo de execução diante dos demais, para todos os experimentos. Os algoritmos com restrições binárias diminuem o universo de soluções do problema e fazem menos multiplicações matriciais durante a execução, portanto, são mais rápidos. WC-FNMTF e FNMTF foram mais rápidos tanto nos experimentos com dados sintéticos quanto nos experimentos com dados do mundo real. O algoritmo ONM3F teve bons resultados com dados sintéticos e o NBVD com dados do mundo real. A depender do problema a ser resolvido, o tempo de execução de um algoritmo pode ser um fator importante de escolha.

[13] Os algoritmos não foram avaliados do ponto de vista de complexidade computacional, apenas sob a ótica de tempo de execução

Tabela 10 – Tempo de execução (medido em segundos) dos algoritmos de coagrupamento para os experimentos com dados sintéticos

Algoritmo	Conjunto de dados sintéticos									Média
	11a	11b	11c	11d	11e	11f	11g	11h	11i	
NBVD	1,365	1,948	1,551	4,494	3,986	1,945	1,918	2,007	2,264	2,387
ONM3F	1,266	0,176	0,123	0,207	0,195	0,177	0,214	0,180	0,202	0,304
ONMTF	1,948	1,840	1,724	1,046	0,667	1,267	1,469	0,842	0,731	1,282
OvNMTF	7,812	17,224	17,913	20,153	25,168	23,002	19,440	40,946	29,021	22,298
FNMTF	0,082	0,087	0,118	0,105	0,096	0,093	0,107	0,086	0,080	0,095
BinOvNMTF	0,284	0,363	0,430	0,773	0,592	0,340	0,342	0,654	0,560	0,482
WC-NMTF	4,700	4,958	5,081	10,262	7,045	4,790	5,250	5,830	6,226	6,016
WC-FNMTF	0,122	0,163	0,218	0,207	0,200	0,200	0,228	0,264	0,247	0,205

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Tabela 11 – Tempo de execução dos algoritmos de coagrupamento para os experimentos com dados do mundo real

Algoritmo	Conjuntos de dados do mundo real			Média
	Binária	TF	TF-IDF	
NBVD	19,215	4,438	2,521	8,725
ONM3F	176,247	59,465	36,484	90,732
ONMTF	44,344	24,947	15,893	28,395
OvNMTF	230,150	15,245	20,990	88,795
FNMTF	10,611	6,784	9,276	8,890
BinOvNMTF	25,393	22,340	28,480	25,404
WC-NMTF	227,757	218,941	220,441	222,380
WC-FNMTF	4,098	4,117	3,680	3,965

Fonte: Waldyr Lourenço de Freitas Junior, 2023

5 Conclusão

Esta dissertação teve como principal objetivo explorar sistematicamente uma série de algoritmos de coagrupamento baseados em fatoração de matrizes, e avaliar os resultados produzidos por eles sob a ótica de interpretabilidade humana. O principal estímulo para esta tarefa decorreu da lacuna encontrada na literatura no que diz respeito à interação humana na análise dos grupos e cogrupos produzidos pelos algoritmos.

Um outro objetivo deste trabalho foi sistematizar formas de extração de informação de grupos dos resultados das fatorações dos diferentes algoritmos e organizar essas informações de modo que pessoas pudessem analisá-las, propondo assim um método adequado de análise qualitativa. Ademais, como parte desta tarefa, o trabalho propôs uma análise do desempenho dos diversos algoritmos de coagrupamento, sob um ponto de vista quantitativo, a fim de fornecer uma análise detalhada e comparativa dos resultados dessa classe de algoritmos.

Diante dessas proposições, algumas questões de pesquisa foram estabelecidas a fim de nortear este trabalho e uma série de experimentos foram realizados para atingir os objetivos supracitados. As questões de pesquisa foram discutidas na seção 1.2 e os experimentos foram apresentados e detalhados no capítulo 4. As análises realizadas levaram em consideração bases de dados sintéticas, como prova de conceito, e uma base de dados do mundo real, como exemplo de aplicabilidade. Os códigos fontes dos algoritmos, os conjuntos de dados, todos os gráficos dos experimentos, as planilhas com as respostas relacionadas à análise qualitativa estão disponibilizados em um repositório público[1].

5.1 Contribuições

Diante do exposto, as principais contribuições deste trabalho são:

- Proposição do algoritmo WC-FNMTF, baseado na ideia do algoritmo WC-NMTF (SALAH; AILEM; NADIF, 2018) e cuja derivação foi baseada nas regras de fatoração de FNMTF (WANG *et al.*, 2011), acompanhado de uma derivação formal das regras de atualização das matrizes U, S, V e Q. A proposição desse algoritmo, embora não fizesse parte dos objetivos deste trabalho, se deu a partir da verificação

[1] https://github.com/waldyrjunior/textCoclustering

da possibilidade de propor uma versão alternativa ao algoritmo WC-NMTF que considerasse restrições binárias, trazendo mais sistematização para os procedimentos comparativos realizados na pesquisa.

- Estabelecimento de uma estrutura de análise dos resultados dos algoritmos sob uma ótica de agrupamento das linhas, colunas, esparsidade e erro de reconstrução, e um comparativo de algoritmos com e sem restrições binárias.
- Proposição de uma estratégia de extração de informação com objetivo de possibilitar a interpretação humana para os resultados produzidos por algoritmos que possuem restrições binárias.
- Apresentação de uma análise qualitativa, com base na interpretação humana, dos resultados produzidos pelos algoritmos de coagrupamento aplicados a dados do tipo texto.

Os resultados obtidos com os experimentos deste trabalho permitiram demonstrar que os algoritmos de coagrupamento produziram resultados melhores, de uma forma geral, que o algoritmo *k-means*, quando implementado sobre as regras de fatoração de matrizes (processo semelhante a NMF). Essa constatação foi possível tanto sob as análises quantitativas quanto qualitativas. Além disso, os algoritmos que possuem restrições binárias produziram resultados melhores que aqueles que não as possuem, para a maioria dos experimentos.

Com relação aos experimentos realizados com interação humana, não foi possível encontrar uma correlação forte entre a capacidade dos algoritmos de agrupar dados (avaliação quantitativa) e a qualidade da informação produzida mediante o resultado do agrupamento (avaliação qualitativa). Por exemplo, o algoritmo FNMTF que apresentou os piores resultados segundo a avaliação qualitativa, atingiu os melhores valores de ARI durante os demais experimentos. O algoritmo que apresentou os melhores resultados segundo a avaliação qualitativa, não atingiu os melhores índices de validação de agrupamento nos demais experimentos. O algoritmo WC-FNMTF, proposto neste trabalho, que em grande parte dos resultados das análises qualitativas atingiu bons resultados, também atingiu bons resultados com os índices de agrupamento.

Uma observação adicional deste trabalho, que não fazia parte do escopo original de análises pretendidas, é que os algoritmos sem restrições binárias possuem tempo de

processamento elevado[2]. Os algoritmos com restrições binárias diminuem o universo de soluções do problema e fazem menos multiplicações matriciais durante a execução, portanto, são mais rápidos. Considerando que os experimentos demonstraram que essa classe de algoritmos (com restrições binárias) produz os melhores resultados, tanto sob uma ótica quantitativa quanto sob uma ótica de interpretabilidade humana, na prática, eles são ótimos candidatos para diferentes problemas.

Embora os principais objetivos deste trabalho tenham sido atingidos, há algumas limitações a serem ponderadas:

- Os algoritmos utilizados nos experimentos nem sempre atingem a convergência almejada, uma vez que a execução finaliza antes de atingir o mínimo global.
- Os experimentos com humanos foram realizados com textos e palavras no idioma inglês e isso pode ter dificultado a realização das tarefas, influenciando assim as respostas dos alunos. Para minimizar esta limitação, foi recomendado que os alunos utilizassem aplicativos de tradução de idioma (não foi realizada nenhuma avaliação de proficiência em inglês com os alunos).
- As formas de avaliação podem não retratar alguns aspectos decorrentes dos resultados dos algoritmos. Na avaliação qualitativa, palavras que mais bem representam os grupos foram utilizadas, porém, poder-se-ia optar pela análise das sentenças em que elas aparecem.
- O corpus utilizado neste trabalho, apesar de robusto e completo no sentido de características dos cenários que era proposto avaliar, possui uma temática complexa de análise (notícias de política).
- Os experimentos realizados neste trabalho e apresentados no capítulo 4 não foram submetidos a testes estatísticos para sustentar a significância entre as diferenças dos resultados obtidos.

A ausência de conjuntos de *benchmark* para avaliar grupos de atributos (grupos de colunas) permanece sendo um desafio geral para o problema de coagrupamento, assim como a subjetividade da temática de interpretabilidade humana. Essas limitações estão além das limitações encontradas neste trabalho.

[2] Este trabalho não propôs realizar uma análise detalhada da complexidade dos algoritmos aqui utilizados

5.2 Trabalhos futuros

Durante a execução deste trabalho, uma série de conclusões e questões foram levantadas e abrem espaço para explorar novos aspectos e realizar pesquisas mais específicas, sobretudo nos aspectos que não foram tratados neste trabalho. Alguns destes aspectos incluem:

- Explorar a avaliação humana dos resultados dos algoritmos em diferentes idiomas, aplicando as tarefas para nativos dos idiomas selecionados. Essa estratégia pode levar a uma maior clareza no sentido de interpretabilidade.
- Avaliar o comportamento dos algoritmos de coagrupamento baseados em fatoração de matrizes em diferentes circunstâncias e tarefas, a fim de revelar mais claramente suas vantagens e desvantagens.
- Projetar um estudo específico para isolar as variáveis de inicialização dos algoritmos, uma vez que eles são sensíveis à inicialização e podem atingir diferentes soluções em cada execução.
- Realizar um estudo mais aprofundado do algoritmo WC-FNMTF, proposto neste trabalho.
- Explorar a influência do hiperparâmetro N (utilizado para controle da esparsidade da matriz PPMI) nos algoritmos WC-NMTF e WC-FNMTF. Uma análise mais aprofundada da influência deste hiperparâmetro no resultado dos algoritmos pode contribuir com a análise qualitativa realizada sobre os resultados produzidos.
- Estudar o efeito da estratégia de múltiplas matrizes ao algoritmo WC-FNMTF, semelhante à estratégia do algoritmo BinOvNMTF. Esse tipo de fatoração poderia oferecer resultados interessantes para as tarefas de agrupamento e coagrupamento.

Referências

ABE, H.; YADOHISA, H. Orthogonal nonnegative matrix tri-factorization based on tweedie distributions. *Advances in Data Analysis and Classification*, v. 13, p. 825–853, October 2019. Citado 6 vezes nas páginas 23, 51, 52, 56, 57 e 60.

AILEM, M.; AGHILES, S.; NADIF, M. Non-negative matrix factorization meets word embedding. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery (ACM), 2017. p. 1081–1084. Citado 9 vezes nas páginas 21, 50, 51, 52, 54, 57, 58, 60 e 93.

ALLAB, K.; LABIOD, L.; NADIF, M. SemiNMF-PCA framework for sparse data co-clustering. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2016. p. 347–356. Citado 8 vezes nas páginas 21, 50, 51, 52, 56, 57, 60 e 93.

ALZAHRANI, S.; CERAN, B.; ALASHRI, S.; RUSTON, S. W.; CORMAN, S. R.; DAVULCU, H. Story forms detection in text through concept-based co-clustering. In: *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*. United States: Institute of Electrical and Electronics Engineers Inc., 2016. p. 258–265. Citado 6 vezes nas páginas 21, 51, 53, 54, 58 e 60.

BRUNIALTI, L. F. *Fatoração de matrizes no problema de coagrupamento com sobreposição de colunas*. Dissertação (Mestrado) — Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, Brasil, 2016. Citado 8 vezes nas páginas 31, 33, 34, 35, 37, 38, 39 e 40.

BRUNIALTI, L. F.; PERES, S. M.; SILVA, V. F. da; LIMA, C. A. de M. The BinOvNMTF algorithm: Overlapping columns co-clustering based on non-negative matrix tri-factorization. In: *Brazilian Conference on Intelligent Systems, BRACIS*. Uberlândia, Brazil: IEEE - Conference Publishing Services, 2017. p. 330–335. Citado 11 vezes nas páginas 21, 26, 39, 51, 56, 58, 60, 70, 71, 75 e 93.

CASALINO, G.; CASTIELLO, C.; BUONO, N. D.; MENCAR, C. A framework for intelligent twitter data analysis with non-negative matrix factorization. *International Journal of Web Information Systems*, Emerald Publishing Limited, v. 14, n. 3, p. 334–356, January 2018. Citado 11 vezes nas páginas 21, 28, 50, 51, 52, 53, 54, 58, 59, 60 e 93.

CASALINO, G.; Del Buono, N.; MENCAR, C. Subtractive clustering for seeding non-negative matrix factorizations. *Information Sciences*, v. 257, p. 369–387, February 2014. ISSN 0020–0255. Citado na página 28.

CHEN, Y.; WANG, L.; DONG, M. Semi-supervised document clustering with simultaneous text representation and categorization. *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Springer Berlin Heidelberg, v. 5781, p. 211–226, September 2009. Citado 7 vezes nas páginas 50, 51, 52, 55, 57, 60 e 93.

CHOO, J.; LEE, C.; REDDY, C. K.; PARK, H. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, IEEE Computer Society, v. 19, n. 12, p. 1992–2001, October 2013. Citado na página 28.

DESGRAUPES, B. Clustering indices. In: . Paris, France: University Paris Ouest Lab Modal'X, 2013. p. 34. Citado na página 45.

DING, C. H. Q.; HE, X. On the equivalence of nonnegative matrix factorization and spectral clustering. In: *Proc. of the 2005 Int. Conf. on Data Mining, Newport Beach, CA, USA*. [S.l.: s.n.], 2005. p. 606–610. Citado 4 vezes nas páginas 30, 31, 60 e 70.

DING, C. H. Q.; LI, T.; PENG, W.; PARK, H. Orthogonal nonnegative matrix tri-factorizations for clustering. In: *Proceedings of the 12th ACM SIGKDD Conference on Knowledge Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, 2006. p. 126–135. Citado 6 vezes nas páginas 21, 26, 33, 34, 35 e 56.

FEBRISSY, M.; SALAH, A.; AILEM, M.; NADIF, M. Improving NMF clustering by leveraging contextual relationships among words. *Neurocomputing*, v. 495, p. 105–117, July 2022. Citado 7 vezes nas páginas 23, 50, 51, 52, 55, 57 e 60.

FREITAS JR., W. L.; PERES, S. M.; SILVA, V. F. da; BRUNIALTI, L. F. OvNMTF algorithm: an overlapping non-negative matrix tri-factorization for coclustering. In: *Proceedings of International Joint Conference on Neural Network*. Glasgow, UK: IEEE World Congress on Computational Intelligence, 2020. Citado 14 vezes nas páginas 21, 23, 24, 26, 34, 51, 52, 53, 56, 58, 60, 70, 71 e 73.

GIBBS, G. R. *Análise de dados qualitativos*. Artmed Editora S.A., 2009. ISBN 9788536321332. Disponível em: ⟨http://bds.unb.br/handle/123456789/739⟩. Citado na página 27.

GIL, A. C. *Métodos e Técnicas de Pesquisa Social*. 6th. ed. [S.l.]: Atlas, 2008. ISBN 9788522451425. Citado na página 25.

GUO, F.; HE, Z. shi; LI, L.; XUAN, J. Unsupervised learning of multi-sense embedding with matrix factorization and sparse soft clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 33, n. 13, December 2019. Citado 5 vezes nas páginas 51, 53, 54, 59 e 60.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: Part I. *SIGMOD Record*, Association for Computing Machinery, New York, NY, USA, v. 31, n. 2, p. 40–45, June 2002. ISSN 0163-5808. Citado na página 25.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Incorporated, 2011. ISBN 0123814790. Citado na página 20.

HARTIGAN, J. A. Direct clustering of a data matrix. *Journal of the American Statistical Association*, American Statistical Association, Taylor & Francis, v. 67, n. 337, p. 123–129, March 1972. Citado 2 vezes nas páginas 20 e 31.

HASSANI, A.; AMIR, I.; MANSOURI, N. Text mining using nonnegative matrix factorization and latent semantic analysis. *Neural Computing and Applications*, Springer-Verlag, v. 33, n. 20, p. 13745–13766, October 2021. Citado 8 vezes nas páginas 50, 51, 52, 53, 55, 57, 58 e 60.

HOFMANN, T. Probabilistic latent semantic analysis. In: *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI1999)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. p. 289–296. Citado na página 21.

HOFMANN, T.; PUZICHA, J.; JORDAN, M. I. Learning from dyadic data. In: *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. Cambridge, MA, USA: MIT Press, 1999. v. 11, p. 466–472. Citado na página 22.

HUANG, S.; XU, Z.; LV, J. Adaptive local structure learning for document co-clustering. *Knowledge-Based Systems*, v. 148, p. 74–84, May 2018. Citado na página 21.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, Springer International Publishing, v. 2, p. 193–218, December 1985. Citado 3 vezes nas páginas 26, 46 e 47.

IBRAHIM, R.; ELBAGOURY, A.; KAMEL, M. S.; KARRAY, F. Tools and approaches for topic detection from twitter streams: Survey. *Knowledge and Information Systems*, Springer-Verlag, v. 54, n. 3, p. 511–539, March 2018. Citado na página 28.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. *ACM Computing Surveys*, Association for Computing Machinery, New York, NY, USA, v. 31, n. 3, p. 264–323, October 1999. Citado 2 vezes nas páginas 20 e 29.

LEE, D. D.; SEUNG, S. H. Learning the parts of objects by nonnegative matrix factorization. *Nature*, v. 401, p. 788–791, October 1999. Citado 8 vezes nas páginas 21, 28, 50, 51, 52, 54, 60 e 93.

LEE, D. D.; SEUNG, S. H. Algorithms for non-negative matrix factorization. In: *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000. v. 13, p. 556–562. Citado na página 54.

LEITÃO, C. F.; PRATES, R. O. A aplicação de métodos qualitativos em computação. In: *Jornadas de Atualização em Informática*. Porto Alegre, RS: Sociedade Brasileira de Computação - SBC, 2017. p. 43–90. Citado na página 27.

LEVY, O.; GOLDBERG, Y. Neural word embedding as implicit matrix factorization. In: *Advances in Neural Information Processing Systems*. Montreal, Canadá: Curran Associates, Inc., 2014. p. 2177–2185. Citado na página 41.

LI, Y.; NGOM, A. The non-negative matrix factorization toolbox for biological data mining. *Source Code for Biology and Medicine*, BioMed Central, v. 8, p. 10, Apr 2013. Citado na página 54.

LIU, Y.; HUA, J.; CHEN, Y. Semantic-constraint graph dual non-negative matrix factorization in text co-clustering. In: *2019 International Conference on Image and Video Processing, and Artificial Intelligence*. Shanghai, China: SPIE, 2019. v. 11321, p. 442–447. Citado 4 vezes nas páginas 23, 51, 54 e 60.

LONG, B.; ZHANG, Z. M.; YU, P. S. Co-clustering by block value decomposition. In: *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, 2005. p. 635–640. Citado 7 vezes nas páginas 20, 21, 26, 28, 32, 33 e 55.

MADEIRA, S. C.; OLIVEIRA, A. L. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, v. 1, n. 1, p. 24–45, August 2004. Citado 2 vezes nas páginas 26 e 61.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008. ISBN 0521865719. Citado na página 82.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, American Statistical Association, Taylor & Francis, v. 66, n. 336, p. 846–850, December 1971. Citado na página 46.

ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 11 1987. Citado 3 vezes nas páginas 26, 47 e 48.

SALAH, A.; AILEM, M.; NADIF, M. Word co-occurrence regularized non-negative matrix tri-factorization for text data co-clustering. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, (AAAI-18)*. New Orleans, Louisiana, USA: AAAI Press, 2018. p. 3992–3999. Citado 17 vezes nas páginas 21, 23, 24, 26, 40, 43, 50, 51, 52, 55, 56, 57, 60, 78, 86, 93 e 110.

SEIDMAN, I. *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. Teachers College Press, 2006. ISBN 9780807746660. Disponível em: ⟨https://books.google.com.br/books?id=pk1Rmq-Y15QC⟩. Citado na página 27.

SHAHID, N.; ILYAS, M. U.; ALOWIBDI, J. S.; ALJOHANI, N. R. Word cloud segmentation for simplified exploration of trending topics on twitter. *IET Software*, IEEE, v. 11, n. 5, p. 214–220, October 2017. Citado 5 vezes nas páginas 51, 53, 54, 58 e 60.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition, Fourth Edition*. 4th. ed. USA: Academic Press, Inc., 2008. ISBN 1597492728. Citado na página 44.

WANG, D.; OGIHARA, M. Finding trendy products from pins. In: *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. USA: IEEE, 2015. p. 428–431. Citado na página 21.

WANG, H.; NIE, F.; HUANG, H.; MAKEDON, F. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Volume Two*. Menlo Park, California: AAAI Press, 2011. p. 1553–1558. Citado 7 vezes nas páginas 21, 26, 37, 38, 60, 93 e 110.

XU, W.; LIU, X.; GONG, Y. Document clustering based on non-negative matrix factorization. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2003. p. 267–273. Citado na página 21.

YEUNG, K. Y.; RUZZO, W. Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper “An empirical study on Principal Component Analysis for clustering gene expression data”. *Science*, v. 17, January 2001. Citado na página 47.

YOO, J.; CHOI, S. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information Processing and Management: an International Journal*, v. 46, p. 559–570, September 2010. ISSN 0306-4573. Citado 15 vezes nas páginas 21, 23, 24, 26, 29, 33, 34, 35, 51, 52, 53, 54, 56, 57 e 60.

Apêndice A – Termo de Consentimento Livre e Esclarecido

Você está sendo convidado(a) a participar como voluntário(a) em uma pesquisa acadêmica intitulada "Estudo com participação de pessoas na análise de resultados de algoritmos de agrupamento de textos", sob responsabilidade do pesquisador Waldyr Lourenço de Freitas Junior, estudante de mestrado do Programa de Pós-graduação em Sistemas de Informação (PPgSI) da Escola de Artes, Ciências e Humanidades (EACH) da Universidade de São Paulo (USP), orientado pela Professora Dra. Sarajane Marques Peres, docente do PPgSI. O objetivo desta pesquisa é coletar dados textuais analisados pelos participantes e documentá-los de forma que sejam utilizados para explicar os resultados de algoritmos de agrupamento e coagrupamento sob uma ótica semântica.

Você realizará algumas tarefas com textos de notícias na língua inglesa.

Os prováveis riscos relacionados e esta pesquisa compreendem cansaço físico e/ou tédio, devido à necessidade de interação constante com o computador. O tempo previsto de sua participação é de 10 a 20 minutos por tarefa e você poderá interromper sua participação sempre que quiser. Sua participação contribuirá para avanços em pesquisa na área de mineração de textos.

Você será esclarecido(a) sobre a pesquisa em qualquer aspecto que desejar e será livre para recusar-se a participar, retirar seu consentimento ou interromper a participação a qualquer momento. A sua participação é voluntária e a sua recusa em participar não irá acarretar qualquer penalidade.

Os pesquisadores irão tratar a sua identidade com respeito e seguirão padrões profissionais de sigilo, assegurando e garantindo o sigilo e confidencialidade dos dados pessoais dos participantes de pesquisa. Seu nome, ou qualquer material que indique a sua participação não será liberado sem a sua permissão. Você não será identificado(a) em nenhuma publicação que possa resultar deste estudo.

Você não terá nenhum gasto e ganho financeiro por participar desta pesquisa.

Qualquer dúvida a respeito da pesquisa, por favor entre em contato com os pesquisadores responsáveis e com o Comitê de Ética em Pesquisa em Seres Humanos (CEP) da EACH/USP, cujos e-mails encontram-se a seguir.

- Waldyr Lourenço de Freitas Junior: waldyrjunior@usp.br; 11 97681 0913.
- Prof. Dra. Sarajane Marques Peres: sarajane@usp.br; 11 98711 0994.

- CEP – EACH/USP: cep-each@usp.br; 11 3091 1046.

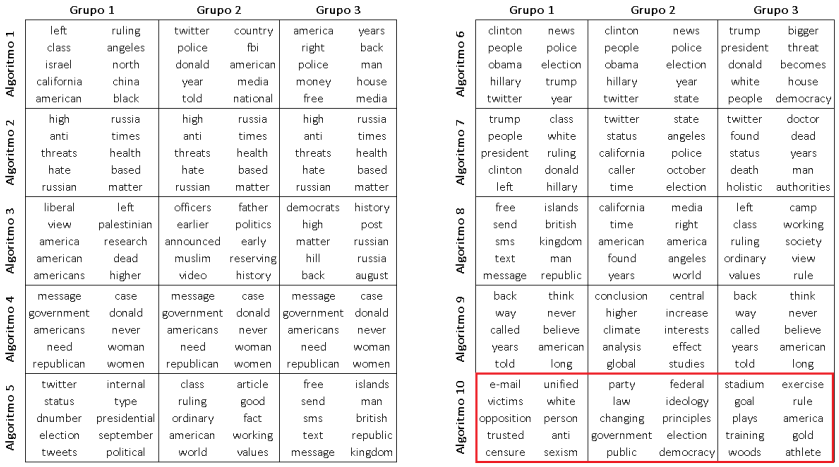
Declaro que fui informado(a) dos objetivos da pesquisa acima de maneira clara e detalhada e esclareci minhas dúvidas. Sei que em qualquer momento poderei solicitar novas informações para motivar minha decisão, se assim o desejar. Os pesquisadores declararam de que todos os dados desta pesquisa serão confidenciais e somente eles terão acesso.

Tenho ciência que o CEP da EACH/USP também poderá ser consultado para eventuais dúvidas/denúncias relacionadas à ética da pesquisa. Este comitê tem a função de implementar as normas e diretrizes regulamentadoras de pesquisas envolvendo seres humanos, aprovadas pelo Conselho Nacional de Saúde.

Declaro que concordo em participar desta pesquisa, voluntariamente, após ter sido devidamente esclarecido.

Apêndice B – Grupos de palavras apresentados aos alunos como parte da atividade 4

Figura 44 – Exemplo do quadro com os grupos de palavras utilizado na atividade 4 para $k = 3$, com destaque para o grupo controlado



	Grupo 1		Grupo 2		Grupo 3	
Algoritmo 1	left	ruling	twitter	country	america	years
	class	angeles	police	fbi	right	back
	israel	north	donald	american	police	man
	california	china	year	media	money	house
	american	black	told	national	free	media
Algoritmo 2	high	russia	high	russia	high	russia
	anti	times	anti	times	anti	times
	threats	health	threats	health	threats	health
	hate	based	hate	based	hate	based
	russian	matter	russian	matter	russian	matter
Algoritmo 3	liberal	left	officers	father	democrats	history
	view	palestinian	earlier	politics	high	post
	america	research	announced	early	matter	russian
	american	dead	muslim	reserving	hill	russia
	americans	higher	video	history	back	august
Algoritmo 4	message	case	message	case	message	case
	government	donald	government	donald	government	donald
	americans	never	americans	never	americans	never
	need	woman	need	woman	need	woman
	republican	women	republican	women	republican	women
Algoritmo 5	twitter	internal	class	article	free	islands
	status	type	ruling	good	send	man
	dnumber	presidential	ordinary	fact	sms	british
	election	september	american	working	text	republic
	tweets	political	world	values	message	kingdom

	Grupo 1		Grupo 2		Grupo 3	
Algoritmo 6	clinton	news	clinton	news	trump	bigger
	people	police	people	police	president	threat
	obama	election	obama	election	donald	becomes
	hillary	trump	hillary	year	white	house
	twitter	year	twitter	state	people	democracy
Algoritmo 7	trump	class	twitter	state	twitter	doctor
	people	white	status	angeles	found	dead
	president	ruling	california	police	status	years
	clinton	donald	caller	october	death	man
	left	hillary	time	election	holistic	authorities
Algoritmo 8	free	islands	california	media	left	camp
	send	british	time	right	class	working
	sms	kingdom	american	america	ruling	society
	text	man	found	angeles	ordinary	view
	message	republic	years	world	values	rule
Algoritmo 9	back	think	conclusion	central	back	think
	way	never	higher	increase	way	never
	called	believe	climate	interests	called	believe
	years	american	analysis	effect	years	american
	told	long	global	studies	told	long
Algoritmo 10	e-mail	unified	party	federal	stadium	exercise
	victims	white	law	ideology	goal	rule
	opposition	person	changing	principles	plays	america
	trusted	anti	government	election	training	gold
	censure	sexism	public	democracy	woods	athlete

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Figura 45 – Exemplo do quadro com os grupos de palavras utilizado na atividade 4 para $k = 4$, com destaque para o grupo controlado

	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
Algoritmo 1	american	think	police	media	class	ordinary	twitter	text
	man	government	twitter	american	ruling	holistic	free	message
	years	women	told	national	found	levels	send	september
	media	states	man	video	gold	death	status	tweets
	political	bill	fbi	think	data	warming	sms	characters
Algoritmo 2	russia	old	russia	old	russia	old	twitter	california
	home	israel	home	israel	home	israel	ruling	page
	media	election	media	election	media	election	left	text
	case	times	case	times	case	times	status	https
	great	country	great	country	great	country	free	message
Algoritmo 3	holistic	good	john	help	text	zuo	storm	control
	gold	far	hollywood	point	send	fallon	think	harvey
	family	party	top	police	free	fake	began	place
	live	important	held	policy	message	fall	event	thursday
	values	warming	bannon	political	sms	falling	become	mccain
Algoritmo 4	george	report	fans	kill	become	inside	nominee	presidential
	campus	bill	skin	killed	health	hate	emails	attacks
	reporter	missile	belly	killer	head	sanders	free	help
	reportedly	campaign	kidnapping	knew	post	harvey	great	host
	reported	talking	cnn	desperate	posted	barack	democracy	need
Algoritmo 5	people	values	found	angeles	twitter	left	trump	hillary
	left	white	california	death	status	september	president	obama
	class	camp	police	caller	trump	presidential	clinton	white
	ruling	working	man	year	election	type	people	american
	ordinary	view	time	holistic	tweets	internal	donald	media
Algoritmo 6	trump	house	twitter	news	clinton	election	twitter	year
	president	america	trump	year	hillary	obama	police	man
	donald	country	people	man	trump	comey	people	august
	people	obama	police	august	fbi	people	obama	right
	white	american	obama	president	president	donald	news	october
Algoritmo 7	clinton	man	twitter	clinton	police	september	trump	hillary
	people	news	august	article	video	year	president	house
	hillary	think	obama	people	state	officers	donald	people
	fbi	right	status	election	officer	children	clinton	obama
	obama	time	october	news	twitter	left	white	campaign
Algoritmo 8	clark	wish	trusted	trump	france	irresponsible	plays	cabinet
	room	vehicles	hours	intentionally	seeking	bringing	wounded	mean
	opposition	skip	star	truth	situation	replied	form	witnessed
	updated	roll	divorce	predators	term	involves	america	nick
	calm	district	efforts	game	mental	increased	main	daughters
Algoritmo 9	player	boys	called	right	called	right	called	right
	color	racists	years	world	years	world	years	world
	sit	progress	back	country	back	country	back	country
	forever	behavior	year	political	year	political	year	political
	athletes	racial	way	including	way	public	way	public
Algoritmo 10	e-mail	unified	party	federal	stadium	exercise	anywhere	regardless
	victims	white	law	ideology	goal	rule	beside	should
	opposition	person	changing	principles	plays	america	particular	able
	trusted	anti	government	election	training	gold	however	small
	censure	sexism	public	democracy	woods	athlete	term	instead

Fonte: Waldyr Lourenço de Freitas Junior, 2023

Apêndice C – Propriedades matemáticas

Este apêndice apresenta algumas propriedades de matrizes e propriedades de somatórios utilizadas para suportar a leitura do capítulo 2 e derivação das regras de atualização do novo algoritmo WC-FNMTF.

Propriedades da adição:

- Comutatividade: $A + B = B + A$;
- Associatividade: $(A + B) + C = A + (B + C)$;
- Elemento neutro da soma: $A + O = A$, $O = [0]_{n \times m}$;
- Elemento simétrico: $A + (-A) = O$; $(A - A = O)$.

Propriedades do produto por um escalar:

- $\alpha(\beta A) = (\alpha\beta)A$;
- $\alpha(A + B) = \alpha A + \alpha B$;
- $(\alpha + \beta)A = \alpha A + \beta A$;
- $1.A = A$.

Propriedades do produto de matrizes:

- Associativa: $(AB)C = A(BC)$;
- Distributiva: $A(B + C) = AB + AC$;
- $(A + B)C = AC + BC$;
- $\alpha(AB) = (\alpha A)B = A(\alpha B)$.

Propriedades da matriz transposta:

- $(A^T)^T = A$;
- $(A + B)^T = A^T + B^T$;
- $(AB)^T = A^T B^T$;
- $(\alpha A)^T = \alpha A^T$, $\alpha \in \mathbb{R}$.

Propriedades do traço:

- $Tr(A + B) = Tr(A) + Tr(B)$;
- $Tr(\alpha A) = \alpha Tr(A)$;

- $Tr(A^T) = Tr(A)$;
- $Tr(AB) = Tr(BA)$.

Propriedades da inversa de uma matriz:

- $(AB)^{-1} = B^{-1}A^{-1}$;
- $(A^{-1})^{-1} = A$;
- $(A^T)^{-1} = (A^{-1})^T$;
- $det(A^{-1}) = \frac{1}{det(A)}$.

Propriedades de somatórios:

- $\sum_{i=1}^{n} b_i = \sum_{j=1}^{n} b_j$;
- $\sum_{i=1}^{n}(a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$;
- $\sum_{i=1}^{n} b_i a_k = a_k \sum_{i=1}^{n} b_i$;
- $\sum_{i=1}^{n} \sum_{j=1}^{m} b_{ij} = \sum_{j=1}^{m} \sum_{i=1}^{n} b_{ij}$.